

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/89458>

Copyright and reuse:

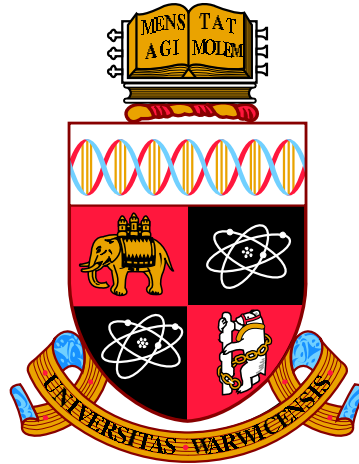
This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk



Variational Bayesian data driven modelling for biomedical systems

by

Yan ZHANG

张雁

Thesis

Submitted to The University of Warwick

for the degree of

Doctor of Philosophy

The School of Engineering

December 2016

THE UNIVERSITY OF
WARWICK

Declaration of Authorship

I, Yan Zhang, declare that this thesis titled, 'Variational Bayesian data driven modelling for biomedical systems' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

“Bayesians address the question everyone is interested in by using assumptions no-one believes, while frequentists use impeccable logic to deal with an issue of no interest to anyone. ”

Louis Lyons

Abstract

Physiological systems are well recognised to be nonlinear, stochastic and complex. In situations when only one time series of a single variable is available, extracting useful information from the dynamic data is crucial to facilitate personalised clinical decisions and deepen the understanding of the underlying mechanisms. This thesis is focused on establishing and validating data-driven models, that incorporate nonlinearity and stochasticity into the model developing framework, to describe a single measurement time series in the field of biomedical engineering. The tasks of model selection and parameter estimation are performed by applying the variational Bayesian method, which has shown great potential as a deterministic alternative to Markov Chain Monte Carlo sampling methods. The free energy, a maximised lower bound of the model evidence, is considered as the main model selection criterion, which penalises the complexity of the model. Several other model selection criteria, alongside the free energy criterion, have been utilised according to the specific requirements of each application. The methodology has been employed to two biomedical applications. For the first application, a nonlinear stochastic second order model has been developed to describe the blood glucose response to food intake for people with and without Diabetes Mellitus (DM). It was found that the glucose dynamics for the people with DM show a higher degree of nonlinearity and a different range of parameter values compared with people without DM. The developed model shows clinical potential of classifying individuals into these two groups, monitoring the effectiveness of the diabetes management, and identifying people with pre-diabetes conditions. For the second application, a linear third order model has been established for the first time to describe post-transplant antibody dynamics after high-risk kidney transplantation. The model was found to have different ranges of parameter values between people with and without acute antibody-mediated rejection (AMR) episodes. The findings may facilitate the formation of an accurate pre-transplant risk profile which predicts AMR and allows the clinician to intervene at a much earlier stage, and therefore improve the outcomes of high-risk kidney transplantation.

Acknowledgements

This thesis would not have been possible without the kind help and support that I have received during the course of my Ph.D from my supervisors, colleagues, family, and friends, without whom I would not be able to overcome the frustrations and difficulties along the way. For this I am eternally indebted and it is a pleasure to have the opportunity to express my gratitude to them all.

My sincerest gratitude goes to my supervisor, Dr. Natasha Khovanova, for her continuous guidance and support throughout four years, with more scheduled and unscheduled meetings than I can count. I am deeply grateful that she offered me this great opportunity to do research in the field of biomedical engineering which I feel passionate about, and introduced me to the inverse problem and the Bayesian inference methods that have intrigued me ever since. During the years, she has offered a tremendous amount of time and energy to advise me and discuss the research problem, and to teach and revise my writing meticulously. Without her, this thesis would never have been completed. I would also like to express my gratitude to my second supervisor, Dr. Neil Evans, and Dr. Igor Khovanov, for all the inspirational ideas, valuable advice and kind help that they have given me over the years. I would like to thank all the collaborators with whom I have worked on my projects. I enjoyed working with them and I have learnt a lot through many meetings and conversations with them.

I am indebted to my good friends, Anup Das, Felicity Kendrick, and Magnus Trägårdh, who have contributed and extended their appreciable help to discuss, proofread, and re-proofread the details in this thesis, which I shall never forget. My thanks also go to all my labmates who are and were in Warwick Engineering in Biomedicine, from whom I have received incredible support and encouragement, their friendship and company have made this Ph.D much more fun and enjoyable than it would have been without them.

Words fail me to express my deepest gratitude to my parents, Zongtao Zhang and Fangming Lin, for their unconditional love, selfless support and incredible patience, and my dog, Xiaohu, for being the best companion for my parents when I am far away in another country.

Contents

Declaration of Authorship	i
Abstract	iii
Acknowledgements	iv
List of Figures	viii
List of Tables	xi
Abbreviations	xii
Symbols	xiv
Disseminations	xvi
Dedication	xix
1 Introduction	1
1.1 Overview	1
1.2 Aim and objectives	8
1.3 Overview of the thesis structure	10
2 Methodology	11
2.1 Model specification	11
2.1.1 Model formulation using ordinary differential equations (ODE) . .	12
2.1.2 Model formulation using stochastic differential equations (SDE) . .	14
2.2 Model selection and parameter estimation	15
2.2.1 Maximum Likelihood and Maximum-a-Posteriori	17
2.2.2 Full Bayes' method	18
2.3 Variational Bayesian scheme	21
2.3.1 Kullback–Leibler divergence and free energy	22
2.3.2 Parameter and state estimation	25
2.3.3 Iterative updating rule for hyperparameters, parameters and the states	26

2.3.4	Free energy decomposition	31
2.4	The criteria for model selection	37
2.5	Priors and hyperpriors	43
2.6	Structural identifiability	47
2.7	Parameter sensitivity analysis	52
2.8	Illustrative example of the use of the VB toolbox	54
2.8.1	The toy model	55
2.8.2	The choice of priors	56
2.8.2.1	Initial default setting of priors and hyperpriors	57
2.8.2.2	Select the mean vector of the prior distribution for the parameters	59
2.8.2.3	Select the variances of prior distributions for the parameters	61
2.8.2.4	Select the hyperpriors for the hyperparameters	63
2.9	Chapter summary	68
3	Post-prandial glucose dynamics	70
3.1	Introduction	71
3.2	Data description	76
3.3	Model and Methods	80
3.3.1	Model formulation	80
3.3.2	Model selection and parameter inference	82
3.3.3	Model selection criteria	85
3.4	Results and Discussions	85
3.4.1	Model selection and parameter inference	85
3.4.2	Structural identifiability and parameter sensitivity	91
3.4.3	Parameter comparison between the groups	93
3.4.3.1	Coefficients of function f_0 and undamped frequency	94
3.4.3.2	Coefficients of function f_1 and damping	94
3.4.3.3	Noise	97
3.4.4	Impulsive force	98
3.4.5	Link between the data-driven model and physiological models. The signs of pre-DM	100
3.5	Conclusions and limitations of the study	103
4	Post-transplant antibody dynamics	105
4.1	Introduction	106
4.2	Data description and visual analysis of dynamic patterns	110
4.3	Models and methods	113
4.3.1	Data fitting and model selection	113
4.3.1.1	Exponential fitting	113
4.3.1.2	Form of the model: linear/nonlinear and stochastic terms	114
4.3.2	Model and parameter identification	115
4.3.3	Model selection criteria	116
4.4	Results and Discussions	117
4.4.1	Model selection	117
4.4.1.1	Comparison of linear models of different orders	118

List of Figures

2.1	Approximation of the parameter distribution of P based on two K-L divergence measures; $q_1(x)$ is based on $KL[Q P]$ and $q_2(x)$ is based on $KL[P Q]$	23
2.2	One iteration cycle of the VB algorithm. In each step, the probability distribution of the i th component q_i ($i = 1, 2, 3, 4, 5$) is optimised to maximise the free energy \mathcal{F} while keeping the probability distributions of the other components fixed. The lower bound of the log-evidence, i.e. the free energy, is guaranteed to increase (or stay unchanged) during each step. The mark * indicates that the probability distribution has been updated.	27
2.3	Illustration of the model M . y_t is the measurement point at time t , and \mathbf{x}_t is the state vector of the system at time t . $\boldsymbol{\theta}$ is the parameter vector of the system equation (2.2). α_η is the precision of the system noise and α_ε is the precision of the measurement noise, u_t is the input of the system at time t	28
2.4	Simulated time series from the toy model.	56
2.5	The correlations between the parameters θ_1 , θ_2 , θ_3 and the initial conditions x_0 , \dot{x}_0 . The correlation between θ_2 and θ_3 is 0.97, between θ_1 and θ_2 is 0.54, and between θ_1 and θ_3 is 0.60.	59
2.6	Free energy and posterior means of θ_3 using different prior means of θ_3	60
2.7	Free energy and posterior means of θ_2 using different prior mean values of θ_2	61
2.8	Free energy and posterior means of θ_1 using different prior means of θ_1	62
2.9	Free energy \mathcal{F} and RMSE_Σ with the variance of the prior distribution for all three parameters $10^j \mathbb{I}_3$ changing from $j = -2$ to $j = 8$	63
2.10	Probability density distribution of the precision of the noise	64
2.11	(a)The value of free energy and (b) RMSE_α for different combinations of the shape a_2 and rate b_2 hyperpriors for the measurement noise. Note that the figure is shown in log-scale (with a base of 10). All the free energy values \mathcal{F} in this graph are negative, so the logarithm of the free energy is calculated by $-\lg(\mathcal{F})$	65
2.12	(a)The value of free energy and (b) RMSE_α for different combinations of the shape a_2 and rate b_2 hyperpriors for the system noise. Note that the figure is shown in log-scale (with a base of 10).	67
3.1	Simple illustration of the glucose - insulin feedback system.	72

3.2	Example subcutaneous glucose time series G of a participant from (a) the control group (b) the T1D group (c) the T2D group. The solid grey curves represent the measured glucose values and the dots are the values used for modelling of single postprandial peaks. The solid and dashed vertical lines correspond to midnight (0 hours) and 6 am respectively. The first several hours of data in Day 1 (to the left from the first solid vertical lines) were excluded from the modelling due to the adjustment period of the CGM system. ‘B’ indicates breakfast, ‘S’ indicates snack, ‘L’ indicates lunch, ‘D’ indicates dinner.	79
3.3	Typical outcome for one peak (the sixth peak in Fig. 3.2 (b)) fitting by four models.	86
3.4	The fitting results are shown for: (a) a peak of a T2D profile; (b) a peak of a T1D profile. The lines are simulated deterministic solutions using the inferred parameters for M_L and M_2	89
3.5	Glucose time series G of a T2D subject. The solid grey curves represent the measured glucose values and the dots are the values used for modelling of single postprandial peaks. The glucose time series corresponding to responses to food intake during breakfasts are indicated by dark black crosses. The solid and dashed vertical lines correspond to midnight (12 am) and 6 am respectively.	91
3.6	The fitting results are shown for three peaks (represent breakfasts) of a T2D profile (as in Fig. 3.5) in three consecutive days from (a) – (c). The free energy value of M_L is denoted as \mathcal{F}_{M_L} , and the free energy value of M_2 is denoted as \mathcal{F}_{M_2}	92
3.7	Boxplots for parameter $\sqrt{\theta_1}/2\pi$ obtained from (a) M_L and (b) M_2 . Note that the denominator 2π is to convert the units from [radian/min] to [min^{-1}]	95
3.8	Boxplots for the damping coefficients $\zeta = \frac{\theta_k}{2 \times \sqrt{\theta_1}}$ obtained from M_L	95
3.9	Boxplots for parameters: (a) θ_k in M_L ; (b) θ_{k_0} in M_2 ; (c) θ_{k_1} in M_2 ; (d) θ_{k_2} in M_2	96
3.10	Boxplots for intensities of system noise I_ϵ in (a) M_L and (c) M_2 , and of measurement I_η noise in (b) M_L and (d) M_2	97
3.11	Boxplots for the initial force parameters F compared between the three groups.	99
3.12	Boxplots for the food impact force F compared (a) between patient No. 15 and the rest of the T2D group; (b) among the subjects in the T2D groups.	100
3.13	The dotted line is the simulated time series from the maximal model for a non-DM case without any signs of DM and the solid line is the deterministic solution using the parameter values inferred from model M_L	101
3.14	Boxplots for (a) θ_{k_0} and (b) θ_1 for all measured peaks fitted by M_L in our cohort of participants. Horizontal lines mark $\theta_{k_0}^{MM}$ and θ_1^{MM} for no signs of DM (upper dashed green line), low insulin sensitivity (middle solid line) and impaired β -cell function (lower dashed pink line) cases.	103
4.1	Measured time series illustrating individual DSA changes in the no-AMR group. Markers correspond to each measurement point. MFI=mean fluorescent intensity.	111

4.2	Measured time series illustrating individual DSA changes in the AMR group. Markers correspond to each measurement point. MFI=mean fluorescence intensity.	112
4.3	Typical fitting results compared among the three models $M_1 - M_3$ for (a) HLA-B60 (case 52) for a patient from the no-AMR group; (b) HLA-DRB3*01 for a patient (case 14) from the AMR group; (c) HLA-A32 for a patient (case 16) from the AMR group; (d) HLA-A2 for a patient (case 17) from the AMR group. The measured values are indicated by circles. .	119
4.4	Boxplot of the difference between the NRMSE of M_2 and NRMSE of M_3 . .	121
4.5	Fitting comparison between M_3S and M_3 for the time series in Fig. 4.3 (b). .	122
4.6	Fitting results compared between the two nonlinear models NM_1 and NM_2 and the linear model M_3 for the time series shown in Fig. 4.3 (c). The measured values are indicated by circles.	124
4.7	NRMSE value of the errors between the observations and the estimated values by M_2 and M_3	125
4.8	Boxplot for the inferred parameters $\theta_0, \theta_1, \theta_2, \theta_3$	127
4.9	Boxplot for the settling values compared between no-AMR and AMR groups .	128
4.10	Phase portraits of the three dimensional system for two DSA time series, (a) from a patient in the AMR group and (b) from a patient in the no-AMR group. The time difference between two consecutive markers is one day.	131
4.11	Boxplot for (a) the larger real part of the eigenvalues; (b) the smaller real part of the eigenvalues; (c) the imaginary part of the eigenvalues $\lambda_{2,3}$; (d) the dissipation rate between the two groups.	132

List of Tables

2.1	Bayes factor compared between models M_1 and M_2	42
2.2	Summary of the parameter settings for the time series simulated by the toy model	56
3.1	Biometric indices, treatment regimens, HbA1c values and corresponding estimated average blood glucose levels of participants.	78
3.2	Four model candidates for fitting	83
3.3	Free energy of M_L , denoted as \mathcal{F}_{M_L} , for different hyperprior settings of the precision of system noise	87
3.4	Free energy of M_2 , denoted as \mathcal{F}_{M_2} , for different hyperprior settings of the precision of system noise	87
3.5	Inference results of the parameters for the example shown in Fig. 3.3 . . .	88
3.6	Summary of peak fitting using M_L and M_2	90
3.7	Summary of the parameter sensitivities for M_L and the range of RMSE with 1% parameter perturbation	93
3.8	Summary of the parameter sensitivities for M_2 and the range of RMSE with 1% parameter perturbation	94
4.1	Summary of the free energy and the NRMSE values for three models of different order corresponding to the four example datasets in Fig. 4.3. . .	120
4.2	Summary of the parameter sensitivities for M_3	127

Abbreviations

AIC	A kaike I nformation C riterion
AICc	C orrected A kaike I nformation C riterion
AiT	A ntibody incompatible T ransplantation
AMR	A ntibody M ediated R ejection
AR	A uto R egressive
AWGN	A dditive W hite G aussian N oise
BIC	B ayesian I nformation C riterion
CGM	C ontinuous G lucose M onitoring
CV	C oefficient of V ariability
DSA	D onor S pecific A ntibody
DM	D iabetes M ellitus
GOF	G oodness O f F it
HLA	H uman L eukocyte A ntigen
K-L	K ullback - L eibler
LS	L east S quare
MAP	M aximal A P osteriori
MCMC	M arkov C hain M onte C arlo
MFI	M ean F luorescence I ntensity
ML	M aximum L ikelihood
MM	M aximal M odel
NRMSE	N ormalised R oot M ean S quare
ODE	O rdinary D ifferential E quation
RMSE	R oot M ean S quare E rror
SDE	S tochastic D ifferential E quation
SISO	S ingle I nput S ingle O utput systems

SMC	S equential M onte C arlo
T1D	T ype 1 D iabetes
T2D	T ype 2 D iabetes
VB	V ariational B ayesian

Symbols

A	State matrix
B	Input matrix
$B_{1,2}$	Bayes factor in favour of model M_1 over model M_2
C	Output matrix
C	Constant
$\delta(t)$	Dirac delta function
\mathcal{E}	Energy term in the calculation for free energy
f_i	Polynomial functions in the system equation
F	Food impact force
\mathcal{F}	Variational free energy, or free energy for short in this thesis
\mathcal{FI}	Fisher information matrix
g	Measurement function in the measurement equation
G	Glucose concentration
$\mathcal{G}a$	Gamma distribution
G_b	Glucose basal level
\mathcal{H}	Shannon entropy, or entropy for short in this thesis
\mathcal{H}	Hessian matrix
\mathcal{I}	Variational energy
I_η	Intensity of system noise
I_ε	Intensity of measurement noise
k	Dimension of the parameter space, or the number of the parameters
$KL(P\ Q)$	Kulback-Leibler divergence from probability density function $q(x)$ to $p(x)$
KL_r	Relative Kulback-Leibler divergence
M	Model
\mathcal{N}	Normal distribution or Gaussian distribution

$p(\cdot)$	Normalised probability density function
$P(\cdot)$	Unnormalised likelihood function
$P(M \mathbf{y})$	Marginal likelihood or model evidence
$P(\mathbf{y} \boldsymbol{\theta}, M)$	Conditional likelihood of obtaining measurement \mathbf{y}
\lg	Logarithm with a base of 10
\log	Natural logarithm
$\log P(\mathbf{y} M)$	Log-evidence of the model
$\log P(\mathbf{y} M, \boldsymbol{\theta})$	Log-likelihood of the model
$q(\cdot)$	Approximated probability density function of $p(\cdot)$
$\langle \cdot \rangle_p$	Expectation of \cdot with respect to the probability distribution p
S	Self-information
SI_{θ_i}	Sensitivity index for the parameter θ_i
u_t	Input at time t
\mathbf{x}	State vector
$\mathbf{x}_t^{(i)}$	i th derivative of the states at time t
\mathbf{y}	Measurement vector
y_t	Measurement data point at time t
$y_t^{(i)}$	i th derivative of the measurement at time t
α_η	Precision of the system noise
α_ε	Precision of the measurement noise
$\gamma(\cdot)$	Gamma function
σ_η	Standard deviation of the system noise
σ_ε	Standard deviation of the measurement noise
Σ	Covariance matrix
η_t	System noise at time t
ε_t	Measurement noise at time t
Θ	All the components that requires iterative updating in the VB scheme
$\boldsymbol{\theta}$	Parameter vector
$\boldsymbol{\theta}^{(i)}$	The i the sample of $\boldsymbol{\theta}$
$\boldsymbol{\phi}$	Parameter vector of the measurement equation
$\psi(\cdot)$	Digamma function

Publications and Disseminations

Results of this research have been disseminated in 6 publications and at 10 conferences as below:

List of Publications

1. Y. Zhang, D. Briggs, D. Lowe, D. Mitchell, S. Daga, N. Krishnan, R. Higgins and N. Khovanova, “A new data-driven model for post-transplant antibody dynamics in high risk kidney transplantation”, doi: 10.1016/j.mbs.2016.04.008, *Mathematical Biosciences*, 2016, in press
2. Y. Zhang, T. A. Holt, and N. Khovanova, “A data driven nonlinear stochastic model for blood glucose dynamics”, *Computer Methods and Programs in Biomedicine*, vol. 125, pp. 18–25, 2015
3. Y. Zhang, D. Lowe, D. Briggs, R. Higgins, and N. Khovanova, “Novel data-driven stochastic model for antibody dynamics in kidney transplantation”, *IFAC–PapersOnLine*, vol. 48, no. 20, pp. 249–254, 2015
4. N. Khovanova, Y. Zhang, and T. A. Holt, “Generalised stochastic model for characterisation of subcutaneous glucose time series”, *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 484–487, June 2014
5. Y. Zhang and N. Khovanova, “A novel stochastic model of postprandial blood glucose time series”, *PGBiomed/ISC–2014: 8th IEEE EMBS International UK & Republic of Ireland Postgraduate Conference on Biomedical Engineering and Medical Physics*, University of Warwick, 2014

6. Y. Zhang, N. Khovanova, and T. A. Holt, “Inference of stochastic nonlinear equations for characterisation and prediction of prandial blood glucose levels”, *ENOC 2014: Proceedings of 8th European Nonlinear Dynamics Conference*, Institute of Mechanics and Mechatronics, Vienna University of Technology, July 2014

List of posters and presentations in conferences

1. Poster presentation (**Best Poster Award**): N. Khovanova, Y. Zhang, D. Lowe, D. Briggs, and R. Higgins, “Dynamic Model for the Evolution of Donor Specific (DSA) Antibodies after HLA Antibody Incompatible (HLAi) Kidney Transplantation”, *BShI 2015: British Society for Histocompatibility and Immunogenetics Annual Conference*, Cambridge, September, 2015
2. Poster presentation: Y. Zhang, D. Lowe, D. Briggs, R. Higgins and N. Khovanova, “Variational Bayesian state-space model for antibody dynamics after kidney transplantation”, *3rd Meeting on Statistics*, Athens, Greece, June, 2015
3. Oral presentation: Y. Zhang, D. Lowe, D. Briggs, R. Higgins, and N. Khovanova, “Novel data driven stochastic model for antibody dynamics in kidney transplantation”, *IFAC2015: 9th IFAC Symposium on Biological and Medical Systems*, Berlin, Germany, August, 2015
4. Oral presentation: Y. Zhang, “Variational Bayesian inference method in stochastic modelling of subcutaneous glucose concentration”, *DINSTOCH: Workshop on Statistical Methods for Dynamical Stochastic Models*, University of Warwick, UK, September, 2014
5. Oral presentation: Y. Zhang, N. Khovanova: “Variational Bayesian analysis of blood glucose time series”, *SCCS2014: the Student Conference on Complexity Science*, Brighton, UK, August 2014
6. Poster presentation: “Inference of stochastic nonlinear equations for characterisation and prediction of prandial blood glucose levels” *ENOC2014: 8th European Nonlinear Dynamics Conference*, Vienna, Austria, July, 2014
7. Member of the organising committee and oral presentation: Y. Zhang, N. Khovanova, “A novel stochastic model of postprandial blood glucose time series”, *the*

8th IEEE EMBS UK & Republic of Ireland Postgraduate Conference on Biomedical Engineering and Medical Physics, University of Warwick, UK, July 2014

8. Poster presentation (**Best Poster Prize Runner-up**): Y. Zhang, “Stochastic modelling of subcutaneous glucose concentration after meals”, $[id]_{Ox}^2$ *inter-disciplinary inter-DTC Student Conference*, Oxford, June, 2014
9. Oral and poster presentation: Y. Zhang, “A data-driven approach for blood glucose modelling”, *Simplifying Assumptions in Models of Complex Systems: Break, Make, Justify*, University of Birmingham, UK, May, 2014
10. Poster presentation (**Best Poster Presentation Award**): Y. Zhang, N. Khovanova, and T. A. Holt, “Stochastic modelling of subcutaneous glucose levels by variational Bayesian inference approach”, Annual Postgraduate symposium, School of Engineering, University of Warwick, March 2014.

Dedicated to my parents

Chapter 1

Introduction

1.1 Overview

Extracting useful information from limited clinical data has always been one of the main focuses of biomedical research. Clinicians need to make decisions – sometimes life-changing ones – for individual cases based on the available data, which can often be sparse and noisy [1]. Clinical data can broadly be split into two categories: one type characterises the static properties such as age, gender, etc., and the other type characterises the dynamic properties of how the physiological system evolves. Whilst the first type of data is usually statistically analysed through data-mining techniques, the second type is often limited to monitoring purposes after certain clinical interventions or treatments. However, the dynamic response of the physiological system after a clinical intervention or treatment carries essential information about possible clinical outcomes, and therefore should be extracted and analysed to help the clinicians provide improved prevention and screening, diagnosis, prognosis, and/or predictions of response to treatment or clinical intervention [2–4]. Various mathematical models have been established to describe, interpret, or predict the dynamic behaviours in clinical data [5–7] to obtain two essential pieces of information: 1) the common patterns shown by different groups of patients and 2) the distinctive features that only belong to an individual. However, when knowledge of the underlying physiological system is limited, it is not straightforward to establish a model that is capable of classifying and recognising patterns in clinical data.

There are two fundamental approaches to developing a mathematical model for describing the dynamics of physiological systems [8], with the purpose of obtaining control of the physiological system to external stimuli. The first approach is based on the understanding of the physiological processes that generate the measured data, which are often referred to as physiological or mechanistic models. Depending on the purpose of the model, this approach can be used to establish models at the molecular, cellular, tissue or organ scale with parameters that have direct physiological interpretations [9]. The process can be time-consuming, may be even impossible in practice, since it requires knowledge at a detailed level about the system's structure and all its parameters. The second approach is focused on describing the measurement data without taking into account explicit knowledge of the physiological processes underneath, which are often referred to as data-driven models. In this case, the system's internal structure is considered as a black box and the only available information about the system is given by its measured inputs and outputs, and the model structure needs to be flexible enough to capture the variations in the data [10]. Data-driven modelling is best implemented when it can be based on the use of inexpensive accessible measurement data to produce parsimonious models [11, 12]. Choice between these two approaches depends on the aim of the research, the level of the understanding of the system, and the availability of the data. The mechanistic approach is better for gaining insight into the working principle of a system, but when the underlying system is not well-understood, establishing a data-driven model may help determine critical characteristics that can provide valuable information for later establishing the mechanistic models [13].

There are several difficulties regarding establishing the first type of model for physiological systems. 1) The most challenging part is to formulate the extraordinary complexity of a physiological system that is the culmination of millions of years of evolution. A system is defined as *complex* when the interactions between the components of the system generate properties that cannot be reduced to its subunits [14]. In physiological systems, new properties have been created as a consequence of the entwined feedback loops that keep humans within the narrow bounds needed for survival [9]. For example, it may be possible for the human body to survive the removal of certain parts by a spontaneous self reorganisation of the system, such as the reshaping of the nervous system following a severe injury, known as brain plasticity. However, a highly intertwined system structure makes it difficult to understand the complex underlying mechanisms. The complexity of

physiological system is handled in nature through multiple feedback mechanisms. This also presents a challenge to system identification due to intrinsic couplings between the variables and strong limitations to the exogenous excitation. 2) It is a formidable task to identify all the parameters in physiological models [15]. Any physiological model that is even approximately realistic will have a large number of parameters [16]. Researchers commonly try to determine the values of these parameters given only one measurement time series, such as the concentration of a certain type of blood cell, the electrical activity of the heart over a period of time (also known as electrocardiography), the human breath rate or the heart rate time series, etc. [17, 18]. In many cases, it is impossible to estimate these parameters, even with perfect data, due to lack of structural identifiability. In cases where the parameter can be estimated in theory, it can still remain statistically challenging, especially in the case where the number of the observation data points in the time series and the number of the parameters are comparable [16]. When certain physiologically based variables in the model are inaccessible, it might be possible to identify these variables in animals where more invasive studies can be conducted [19]. However the translation between human and animal parameters is not as simple as linear scaling [20]. 3) Another important problem is the limited data available for personalised modelling. In clinical applications, individualised models are required for personalised patient care, but many physiological models use a single set of parameters representing the ‘average person’ [16]. However, reliable data acquired through minimally invasive techniques for each patient still remain scarce [21]. 4) The lack of a generalised acceptable model for different physiological processes is another issue, since the nature and structure of underlying processes varies significantly from system to system [22].

The second category - data driven dynamic modelling - does not require a complete understanding of the underlying physiological system and therefore does not have the difficulties that are associated with the physiologically based models. Data-driven models can be categorised into parametric and non-parametric models. A parametric model describes the system using a limited number of characteristic quantities – the parameters of the model – while a non-parametric model determines the model structure directly from the measurements. The term non-parametric does not imply that the model lacks parameters. Instead, it means that the number and the nature of the parameters are flexible and not fixed in advance, leading to less structural interpretability compared to parametric models. However, within parametric models, the information about the

system is captured by a relatively small number of parameters, compared with non-parametric models. Therefore, the physical insights and concentration of information per parameter is more substantial for parametric models than for non-parametric models [10]. This thesis will focus on parametric models, which seek quantitative descriptions of physiological systems based on input-output information derived from the experimental data. They are mathematical descriptions of data, with only implicit correspondence to the underlying physiology. Non-parametric models, such as Volterra models [23], will not be considered in this thesis. Within the data-driven modelling paradigm, one of the most common approaches is autoregressive models. An autoregressive (AR) model describes the output of a time-varying process by a linear combination of its past values and certain stochastic terms representing noise. The advantage of AR models is that the future output can be easily predicted by only considering the previous values. However, most AR models have no structural interpretation; the identification and estimation of the parameters can also be seriously distorted by measurement outliers.

Data-driven models based on differential equations are often overlooked due to the difficulties of selecting the appropriate form to account for the nonlinearity and stochasticity of the system. Furthermore, the parameter estimation of nonlinear continuous-time models is not a trivial task, and can be computationally involved due to the calculation of time-derivatives or integrals of sophisticated nonlinear functions. However, differential equation based models have the following inherent advantages: 1) the estimated model is defined by a unique set of parameters that are not dependent on the sampling interval, which allows extrapolation to explore and predict in the region that is not included in the data. 2) It can handle irregularity in the sampled data better than the difference models [24]. 3) A differential equation can be approximated by a difference equation with a higher order, and therefore, a model based on differential equations allows a more parsimonious representation compared to difference models [25]. 4) Differential equations can easily accommodate certain prior information into the model by specific initial conditions. 5) The most fundamental reason for using differential equations is that the behaviour of physiological systems constantly evolves with time. The underlying physiological process is continuous by nature (without considering molecular or smaller scales), even though the measurement time series are discrete [26]. Therefore, modelling the continuous process by a differential model is more appropriate than a difference model.

One fundamental property of a physiological system is its nonlinearity. It is often possible to use linear models to approximate nonlinear systems, which is an attractive idea because linear models are well established and easy to interpret. It requires much less effort to build linear models compared with nonlinear models. However, linear equations can only lead to exponentially decaying/growing or (damped) periodically oscillating solutions. For linear modelling, all irregular behaviour of the system has to be attributed to some random external input to the system [17]. However, input is not the only source of the irregularities in a system's output: a small intervention in a nonlinear system, such as a disturbance in the initial conditions can have unexpected outcomes which cannot be simply explained by linear models [14]. A linear approximation to a nonlinear system is only valid for a given input range. On the other hand, nonlinear modelling is less straightforward and far less well understood than its linear counterpart. There is no general nonlinear parametric model framework in the system identification literature [10]. Recognising the nonlinear behaviour and formulating the nonlinear differential equation involves a series of trial-and-error processes with a wide range of nonlinear forms to select from, especially when the measured time series are corrupted with noise. This thesis presents methodology for identification of the structure of the nonlinear mathematical models from the available measured input-output data for several important medical applications, and analysis of both linear and nonlinear behaviours of the transient responses of corresponding dynamic physiological systems.

Another important property of a physiological system is its stochasticity. The state of the stochastic system can only be predicted probabilistically, whereas the outcome from a deterministic system can be reproduced as long as the input stays the same [27]. The uncertainty in the model originates from the action of a very large number of factors or 'degrees of freedom'. Stochasticity can be introduced from external stochastic disturbance, such as environmental influences, to intrinsic regulatory responses towards the disturbance [28], but it is not realistic to model such high-dimensional dynamics. Thus, deterministic models disregard the stochastic aspects of physiological systems. Stochastic models couple their deterministic equations to 'noise' which mimics the perpetual action of many unconsidered variables in the system. Noise is an essential part of the physiological system that should not be neglected. A small amount of noise can have an important role in physiological systems (e.g., [29–31]). For example, the normal human heartbeat fluctuates in a complex stochastic manner [32]; the heartbeat time series

from patients with high risk of sudden death showed reduced stochastic property. By accounting for the intrinsic unpredictability of the system, a stochastic model provides a more realistic view of the underlying process. The sources of uncertainty are usually formulated in stochastic models as two types of noise. The first type corresponds to the uncertainties in the observations, which is modelled as the measurement noise. The second is the imperfection of the model, which is modelled as the system noise [33]. Measurement noise usually only has a blurring effect on the observation, and does not influence the evolution of the system. However, when the system noise interacts with the dynamic variables in nonlinear systems, it can lead to effects such as transitions between the stable states. Since it can dramatically modify the deterministic dynamics, the system noise should not be neglected during the modelling procedure, especially in the case of nonlinear systems. Stochastic models come at a price as they are more computationally demanding than deterministic models, and considerably more difficult to fit to experimental data [34]. The latest advances in statistical inference methods make building such stochastic nonlinear models feasible [35, 36], and such stochastic systems are the subject of our investigations.

As stated above, relevant data that reveal the dynamics of the underlying system, especially in physiological systems, can be limited. A reliable method is needed to extract the maximal information from the limited data to provide two essential pieces of information: 1) the estimates for model parameter values; 2) how well the model describes the data. The process of estimating the parameter values in the model is usually referred to as inference and the evaluation of the model is usually referred to as model selection. In certain fields of engineering, such as electronic engineering, mechanical engineering and systems engineering, the interactions between the input and output data have been well studied since more powerful computing and electronic equipment has made measurement collection easy and affordable. In the biomedical field, however, model selection and parameter estimation remain challenging considering the limited input-output relationships that can be observed in physiological systems. When the measurements are disturbed by noise, the distinguishability between different models is reduced, leading to an uncertainty in the final selection of the model. Therefore, it is paramount to choose an appropriate method of model development to best exploit limited clinical data.

The techniques for model selection and parameter estimation can be divided into two major categories: the ‘classical’ or frequentist methods and the probabilistic or Bayesian

methods [37, 38]. These two categories of methods have several philosophical differences. First, the frequentist methods calculate the probability of obtaining the measurement time series assuming that the model is true, whereas the Bayesian methods calculate the probability of the model being true given the time series. Take the comparison of two nested models, where one model is the reduced model of the other, as an example. The most common technique in frequentist methods is hypothesis testing based on the p-values. Treating the reduced model as the ‘null’ hypothesis, the obtained p-value represents the probability of obtaining the measurements assuming the null model is true. The test is an all-or-nothing proposition for rejecting the null hypothesis, without providing any information about the other model [39]. The Bayesian approach, on the other hand, calculates the probability of either model being true based on the data. It provides a more realistic view [40] since we are more interested to know if the model is more probable rather than if the data are more probable. Second, the frequentist method makes a point estimate of the parameter and compares models based on the exact parameter values inferred; whereas the Bayesian inference method expresses the parameters as probability distributions, and the uncertainties in the parameter values are accounted for during the model comparison [41]. As the model is stochastic with the purpose of capturing the uncertainty of the system, a probabilistic inference method fits better with such a purpose [42]. Furthermore, in Bayesian statistics, a prior belief of the parameter distribution is quantified by a probability distribution, and this belief gets updated based on the likelihood of observing the data. Finally the posterior belief of the parameter distribution takes into account the prior information and the support from the data [42]. In the classical methods, there is no such option of including preliminary information about the data. Therefore, based on the above properties and keeping in mind the complexity and stochasticity of the underlying biomedical systems, Bayesian methods will be considered for model selection and parameter identification.

Probabilistic models can be computationally difficult to implement, especially when a large number of parameters need to be estimated and the distributions of the parameters are not in standard forms (or intractable); there are two ways of addressing this issue. The first is to use stochastic sampling methods such as Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) to sample from the unknown distribution. Thanks to the development of these stochastic sampling methods that are capable of simulating high-dimensional distributions of parameters, the Bayesian approach has

gained popularity in parameter inference ever since the 1990s [42, 43]. However, standard MCMC algorithms such as the Metropolis-Hastings algorithm [44] and the Gibbs algorithm [45] cannot provide a quantitative measure for model comparison between model candidates, and therefore, an additional step for model selection is required. In addition, SMC, MCMC, and related sampling methods require large computational power, which is undesirable. The second way is to approximate the intractable distribution rather than sampling from it, and one of the methods that has been well developed in statistical physics is called the Variational Bayesian (VB) method. The VB method breaks down the task of inferring all the parameters into manageable subsets and learns the value of parameters by iteratively optimising one subset whilst keeping the rest fixed [41]. With no need of stochastic sampling, the VB method provides measures of uncertainty for any point estimates for the parameters with relatively low computational cost. This method has been exploited in parameter inference for graphical models amongst the machine learning community, however, there have been relatively few studies from other potential fields such as biomedical research [42]. Stochastic sampling methods such as MCMC might still remain the dominant method in the field of Bayesian inference, but the purpose of this thesis is to show that the VB method can be successfully applied to biological system identification and can yield robust dynamic models capable of capturing essential properties of such biomedical systems from limited data.

1.2 Aim and objectives

The aim of this thesis is to develop and validate nonlinear stochastic data-driven models that describe the transient responses of the underlying physiological systems through one-dimensional clinical measurement time series. This thesis investigates two clinical applications by applying the Variational Bayesian method to identify and select the model with the appropriate degree of complexity, based on the availability and the precision of the measurement data for each application.

In the first application, the aim is to construct a generalised data-driven model of transient glucose responses to the food intake of subjects with and without diabetes that takes into account the complexity, nonlinearity and stochasticity of the underlying glucose regulatory system. The novelty of this model lies in its concise and parsimonious

form adjusted to each food intake, whilst still retaining an ability to generalise over glucose response behaviours seen in different individuals. Maintaining glycaemic stability is one of the primary goals in diabetes management for people with or in progression towards diabetes. For people that are prone to glucose variability, a model that can accurately describe the glucose responses to each food intake can help to monitor and improve their control over diabetes, leading to a healthier life. For diabetic patients who need insulin injections or medications before each food intake, the model facilitates an informed estimation of the insulin and medication needed, tailored to each meal at specific time of the day.

In the second application, the aim is to build an individualised data-driven model for the first time to describe the post-transplant antibody dynamics after high risk kidney transplantation. The understanding of post-transplant antibody behaviours is still in its early stages and the mechanisms controlling the antibody dynamics are not well understood. Seizing the opportunity opened up by the recently developed technique of measuring the antibody levels, a novel mathematical model is constructed in this thesis to extract information from the limited sparse data and to provide insights with regard to better controlling the antibody response in the early post-transplant stage. The establishment of the data-driven model can also provide information about the structure of the physiological system, and therefore lay the foundation for future physiologically based models.

The objectives of the thesis are listed as follows:

- 1) Identify the dynamic features in the transient response of the underlying physiological system from a single measurement time series. Construct model candidates with different levels of complexity that can describe the identified dynamic features.
- 2) Apply the variational Bayesian method to infer the parameters of the model candidates and select the best model with the appropriate degrees of complexity in terms of nonlinearity and stochasticity. Develop appropriate model selection criteria for both applications.
- 3) Perform structural identifiability and parameter sensitivity analysis to assess the reproducibility and robustness of the model.

- 4) Gain clinical insights from the selected models and the inferred parameters, with the aim of improving patient management and treatment in both applications.

1.3 Overview of the thesis structure

The structure of the thesis is as follows: Chapter 1 gives the background information about modelling physiological systems. It outlines the aim and objectives of this thesis. Chapter 2 describes the main methodology of model specification, parameter estimation, model selection, structural identifiability, and parameter sensitivity. Chapter 3 focuses on the modelling of post-prandial glucose dynamics, and Chapter 4 focuses on the modelling of post-transplant antibody dynamics. Both chapters apply the methods described in Chapter 2. The novelty, the main discoveries, and the future directions of this research are presented in Chapter 5.

Chapter 2

Methodology

This chapter describes the underlying methodology for model specification, parameter estimation, model selection, structural identifiability, and parameter sensitivity. It introduces the *Variational Bayesian* (VB) method which is adapted for the applications given in Chapters 3 and 4. This chapter is divided into eight sections. Section 2.1 describes the procedure for specifying the form of the models in this thesis. Section 2.2 discusses commonly used parameter inference and model selection methods. Section 2.3 gives details of the VB method for parameter estimation. Section 2.4 provides a comparison between several model selection criteria and introduces the free energy criterion. Section 2.5 discusses the effect of the choice of the prior distributions on the parameters of the model. Sections 2.6 and 2.7 present several techniques that are used for parameter identifiability and sensitivity analyses respectively.

2.1 Model specification

Model development in data-driven modelling usually focuses on parameter estimation and model selection [46–48]; whereas relatively little attention is given to the other crucial part of the modelling procedure – specifying the appropriate form of the model that is capable of adequately describing the observed data is paramount. In the previous chapter, the advantages of using differential equations, while accounting for nonlinearities and stochasticities in describing dynamic clinical data, were established. The next

step is to formulate the model using differential equations and incorporate the nonlinear and stochastic features into the model.

2.1.1 Model formulation using ordinary differential equations (ODE)

A dynamical system is described by three components: the state space variables, time, and the law of evolution in time. The state space variables are physical variables of the dynamical system that contain all the information needed for evolution of the states. Each state space variable corresponds to the coordinate axes of the *state space*. When assembled as a vector, the state variables form the state vector, and each possible state of the system corresponds to one point within the state space. The second component of a dynamical system, time, can be treated as a discrete or continuous variable. Empirical measurements are taken at discrete time points – typically at sequential integer values, and the time interval between two sequential integers is the time ‘unit’. Continuous time, in contrast, is typically applied to variables that are related to time by functions. The third component of the dynamical system, the law of evolution, is a rule that transforms one point in the state space, representing the state of the system ‘now’, into another point, representing the state of the system one time unit ‘later’. The state of the system starts at certain initial conditions, evolves with or without external inputs, and generates outputs. As described in Chapter 1, the two biomedical processes that have been investigated for this thesis have a common feature: only one measurement time series, denoted as \mathbf{y} ($\mathbf{y} = \{y_1, y_2, \dots, y_T\}$ where y_T is the last measurement at time $t = T$), is available, and only one external input, denoted as \mathbf{u} ($\mathbf{u} = \{u_1, u_2, \dots, u_{T'}\}$ where $u_{T'}$ is the last input at time $t = T'$, note that T' does not necessarily equal T), is given to the system. Therefore, the theoretical description is restricted to single input single output systems (SISO).

A dynamic model can be written in two forms: the input-output form and the state-space form. Both forms essentially carry the same information about the system dynamics, but are applied in different situations [49]. The ‘input-output’ dynamic equation relating the input \mathbf{u} (the input at time t is denoted as u_t) to the output of the system \mathbf{y} (the

measurement at time t is denoted as y) is as follows:

$$\frac{d^n x_t}{dt^n} + \sum_{i=1}^{n-1} f_i(x_t, \boldsymbol{\theta}_i) \frac{d^i x_t}{dt^i} + f_0(x_t, \boldsymbol{\theta}_0) = u_t \quad (2.1a)$$

$$y_t = g(x_t, \boldsymbol{\phi}) \quad (2.1b)$$

In this thesis letters in **bold** font represent vectors. The equation (2.1a) is the dynamic equation of n th order, and (2.1b) is the measurement equation. The equation (2.1a) can also be written in state-space form where the evolution of each state variable is described

by a first-order differential equation, where $\mathbf{x}_t = \begin{pmatrix} x_t \\ \dot{x}_t \\ \vdots \\ x_t^{(n-1)} \end{pmatrix}$ is the state vector at time t .

A set of n initial conditions must be known in order to solve the equations for a given input \mathbf{u} . f_i ($i = 0, 1, \dots, n-1$) and g are the functions of the dynamic and measurement equations respectively, and they can be linear or nonlinear. $\boldsymbol{\theta}_i$ ($i = 0, 1, \dots, n-1$) are the vectors of parameters in (2.1a) and $\boldsymbol{\phi}$ is the vector of parameters in (2.1b). In this thesis, f_i and g_i are assumed to be smooth and continuously differentiable to guarantee the existence and uniqueness of the solution based on the Picard–Lindelöf theorem [50].

To describe a variety of dynamic responses to the inputs \mathbf{u} , a generic form of the dynamic equation is needed, i.e. the functional forms of f_i ($i = 0, 1, \dots, n-1$) need to be determined. As discussed in Chapter 1, a linear form is easy to implement and linear systems are well understood, but a dynamic equation with linear f_i can only describe limited behaviours of the time series. Nonlinear forms, on the other hand, can be chosen from a huge variety of functions, such as polynomial, logarithmic, trigonometric functions etc. Among these forms, polynomial forms, such as Taylor series, have a wide range of nonlinear solutions, and are computationally easy to differentiate and integrate which simplify the procedure for parameter estimation. For the applications given in Chapters 3 and 4, exponential decay features can be identified by visual examination of the measurement time series. Such features agree with the characteristics given by the solutions of the differential equation with the f_i in polynomial forms. In addition, the linear form of a polynomial function is a special case of its nonlinear form (where the polynomial terms with orders larger than one are zeros), both linear and nonlinear behaviours in the time series from different subjects can be unified using a generalised dynamic equation by

using a polynomial form. Therefore, the polynomial form is selected for f_i as a starting point to explore the nonlinear behaviours in the data.

The next task is to identify the polynomial terms to capture the key dynamic patterns expressed in the given data. Linear differential equations are considered first. If the fitting of a model to data is unsatisfactory, higher order polynomial terms will be added to the f_i . As the order of the polynomial terms increases, more varieties of the dynamic patterns can be expressed by the equation; however, a high order f_i may lead to overfitting, and a large number of parameters may cause identifiability issues. The optimal number of the polynomial terms is that at which the benefit gained from increasing the order compensates for the risks associated with increased model complexity. The details of the choice of the functional forms for the two biomedical applications are further explained and justified in Chapters 3 and 4.

2.1.2 Model formulation using stochastic differential equations (SDE)

A system may have one or more sources of noise coupled in one or more ways, and the literature on the different ways that noise can be incorporated in the model – additively or multiplicatively – is abundant [51]. As stated in Chapter 1, noise can be categorised into system noise and measurement noise. System noise disturbs the states, influencing how the system evolves; while measurement noise is introduced into the system when the states are being measured, and therefore does not perturb the evolution of the dynamical system. The analysis of either form of noise is always coupled with the system states because neither the measurement noise nor the system noise is measurable independently. With system and measurement noise considered separately, the ODE in (2.1a) can be written as a *Stochastic Differential Equation* (SDE) in the form of a Langevin equation [52], and the input-output model in (2.1a) and (2.1b) can be written as the system equation and the measurement equation as follows:

$$\begin{cases} \frac{d^n x_t}{dt^n} + \sum_{i=1}^{n-1} f_i(x_t, \theta_i) \frac{d^i x_t}{dt^i} + f_0(x_t, \theta_0) = u_t + \eta_t \\ y_t = x_t + \varepsilon_t \end{cases} \quad (2.2)$$

where η_t corresponds to the system noise, ε_t corresponds to the measurement noise. Additive noise does not depend on the states of the system, and therefore takes a simpler form than multiplicative noise which depends on the states of the system. With

no preliminary information about the noise available and to reduce model complexity, additive noise is selected for the applications considered in this thesis. Both of the noise terms η_t and ε_t are modelled as *Additive White Gaussian Noise* (AWGN) [51], which is a standard noise model to mimic the effect of many random processes. ‘White’ indicates that the noise has no correlation in time, i.e. the noise has no memory. ‘Gaussian’ indicates that the noise intensity is normally distributed: $\eta_t \sim \mathcal{N}(0, I_\eta)$, $\varepsilon_t \sim \mathcal{N}(0, I_\varepsilon)$. The mean values of both noise terms are zero. I_η and I_ε are the intensities of the system and measurement noise respectively. They are equivalent to the variances, σ_η^2 and σ_ε^2 , of the corresponding noise. The precisions of the system and measurement noise, which are the inverse of the variances, are denoted as $\alpha_\eta = 1/\sigma_\eta^2$ and $\alpha_\varepsilon = 1/\sigma_\varepsilon^2$. The noise precisions α_η and α_ε , considered as the parameters of the stochastic part of the model, together with the deterministic system parameters, are aimed to be inferred from the noisy measurement time series using the inference method described in Section 2.3.

2.2 Model selection and parameter estimation

In Section 2.1, a model structure based on SDEs with polynomial functions was selected, which leads to multiple model candidates which differ in the number of polynomial terms. The task of choosing the best model among these candidates is referred to as *model selection* in this thesis. Determining the specific values of model parameters that describe the data is referred to as *parameter estimation* or *inference*.

In classical statistical methods, the parameters of each model need to be estimated before model selection. The model parameters are usually estimated by minimising a cost function which measures a ‘*goodness of fit*’ (GOF) for a model. GOF is evaluated by analysing the differences between observed values and the values expected under the model in question. The differences are referred to as ‘*residuals*’. The most widely used method is to minimise a cost function of the sum of the squared residuals, named *least squares* (LS) estimation [53]. Except for constructing a cost function, there are other ways to measure GOF, among which the most popular way is the *maximum likelihood* (ML) method [54]. The ML method looks for the optimised parameters by maximising the likelihood of obtaining the measurement time series \mathbf{y} given the parameters $\boldsymbol{\theta}$ of the model M , (the likelihood is denoted by $P(\mathbf{y}|\boldsymbol{\theta}, M)$). Both LS and ML estimation optimises the GOF for each model; however, a model with a better GOF value (a smaller

cost or a higher likelihood) does not guarantee a ‘better’ model. With a sufficiently complex model, parameters can be found to fit the observed data with a high level of precision, but the model might have been ‘*overfitted*’ to the data by erroneously fitting the noise as well. Overfitted models tend to be sensitive towards small fluctuations in the measurements, which can lead to spurious predictions. Therefore, a penalty term for model complexity is often introduced to the cost or the likelihood function (such as the AIC and BIC criteria described in Section 2.4 in detail), along with cross-validating techniques [55], to check for model overfitting. The choice of this penalty term is essential for the model selection task. A penalty term that is too large can result in choosing an *underfitted* model. Underfitted models often fail to reproduce important features in the experimental data, which introduce approximation errors into the model — known as *bias*. A model is considered underfitted if there are serial correlations between the residuals [56]. The model with appropriate order and structure needs to seek a balance between overfitting and underfitting.

Compared with these classical approaches, *Bayesian approaches* have emerged as a more effective and informative alternative in the tasks of parameter estimation and model selection where the tasks can be done simultaneously without the need to choose an appropriate penalty term. Bayesian approaches interpret ‘*probability*’ as a quantity that represents a state of knowledge instead of a frequency of a certain event happening [57], and the states of knowledge get updated when more information is given, known as the *Bayes’ rule* introduced by Cournot (1843). Applying Bayes’ rule to model selection, a preference over several models $p(M)$ before accounting for any data is defined as the *prior*. Given the data \mathbf{y} , the conditional probability $p(M|\mathbf{y})$ of the model M being true is defined as the *posterior* distribution, and can be described in mathematical terms as follows:

$$p(M|\mathbf{y}) = \frac{P(\mathbf{y}|M) \times p(M)}{P(\mathbf{y})} \quad (2.3)$$

where $p(\cdot)$ represents a probability density function (normalised function which integrates to one) and $P(\cdot)$ represents a likelihood function which is not necessarily normalised.

The model with the largest $p(M|\mathbf{y})$ among the model candidates is considered the most probable model. When there is no prior preference for any model, $p(M|\mathbf{y})$ is determined by the likelihood function $P(\mathbf{y}|M)$, defined as the *marginal likelihood*. To calculate

$P(\mathbf{y}|M)$, the parameters of the model need to be estimated. In the next two sections, three parameter estimation methods — *maximum likelihood* (classical method), *maximum-a-posteriori* (a bridge between the classical and the full Bayes' method), and the *full Bayes'* methods — are presented.

2.2.1 Maximum Likelihood and Maximum-a-Posteriori

As suggested by the name of the method, *Maximum Likelihood* (ML) looks for the most probable parameter values $\hat{\boldsymbol{\theta}}_{ML}$ by maximising the likelihood function $P(\mathbf{y}|\boldsymbol{\theta}, M)$, which is the likelihood of obtaining \mathbf{y} given the parameter $\boldsymbol{\theta}$ and the model M , as follows:

$$\hat{\boldsymbol{\theta}}_{ML} = \arg \max_{\boldsymbol{\theta}} P(\mathbf{y}|\boldsymbol{\theta}, M) \quad (2.4)$$

The ML parameter estimation method is widely used due to its simplicity. However, under some circumstances, estimating the most probable parameter only based on data can be misleading. For example, assume model M includes a parameter which represents the probability of having an earthquake in city A with no previous record of an earthquake. The most probable value of the parameter is zero based on the record, which would underestimate the risk of the earthquake, especially if it is known that the city is located right on top of a tectonic plate boundary. Such information – known before taking the data into account – is the prior for the parameter. The method that incorporates the prior information into the ML method is the *Maximum-a-Posteriori* (MAP) method, which can be formulated as follows:

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|M)P(\mathbf{y}|\boldsymbol{\theta}, M) \quad (2.5)$$

where $p(\boldsymbol{\theta}|M)$ is the prior of the parameters given the model, and $P(\mathbf{y}|\boldsymbol{\theta}, M)$, the same as the term in (2.4), is the likelihood of obtaining \mathbf{y} given the parameter $\boldsymbol{\theta}$ and the model M .

The MAP method, just like the ML method, only estimates the mode of the posterior distribution of the parameters. In situations where the confidence level of the estimated parameter is of interest, both the ML and MAP methods are not sufficient and the full Bayes' method is needed.

2.2.2 Full Bayes' method

The full Bayes' method provides the probability distribution of the parameters instead of point estimations of the parameters. Given the prior distribution of the parameters, $p(\boldsymbol{\theta}|M)$, the posterior distribution of the parameters can be obtained by normalising the right hand side of (2.5):

$$p(\boldsymbol{\theta}|\mathbf{y}, M) = \frac{p(\boldsymbol{\theta}|M)P(\mathbf{y}|\boldsymbol{\theta}, M)}{P(\mathbf{y}|M)} \quad (2.6)$$

The prior distribution $p(\boldsymbol{\theta}|M)$ can have a big influence on the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}, M)$ (discussed in detail in Section 2.5). The posterior distribution of the parameters $p(\boldsymbol{\theta}|\mathbf{y}, M)$ is obtained by updating the prior belief $p(\boldsymbol{\theta}|M)$ based on data \mathbf{y} ; therefore, it contains information from both the prior $p(\boldsymbol{\theta}|M)$ and the likelihood of obtaining the data given the parameters for a given model $P(\mathbf{y}|\boldsymbol{\theta}, M)$.

The denominator of (2.6) — the marginal likelihood $P(\mathbf{y}|M)$ — can be obtained by integrating over the parameter space as follows:

$$P(\mathbf{y}|M) = \int P(\mathbf{y}, \boldsymbol{\theta}|M) d\boldsymbol{\theta} = \int P(\mathbf{y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M) d\boldsymbol{\theta} = \langle P(\mathbf{y}|\boldsymbol{\theta}, M) \rangle_{p(\boldsymbol{\theta}|M)} \quad (2.7)$$

where $\langle \cdot \rangle_p$ denotes the expectation with respect to the probability density function $p(\boldsymbol{\theta}|M)$ in the subscript. As the marginal likelihood is the normalisation constant of the posterior distribution of the parameters, it is obtained as a by-product of the parameter distribution estimation. As stated in the last paragraph in Section 2.2, the model with the largest value of marginal likelihood is chosen to be the most probable model; therefore, the task of model selection is achieved simultaneously with the task of parameter estimation by using the full Bayes' method.

An important principle for model selection is called the *principle of parsimony*, which states a preference for the simplest possible model that fits the data [48]. The marginal likelihood value intrinsically obeys this principle, because it accounts for the model complexity regarding the dimension of the parameter space by integrating the likelihood function $P(\mathbf{y}|\boldsymbol{\theta}, M)$ over the parameter space. Intuitively, a more complex model with a larger number of parameters can describe more varieties of data compared to a simpler model. However, in situations where both models describe the data equally well, the simpler model is preferable, because the combination of the parameters that gives the

best fit for the simpler model is more likely to occur in a low-dimensional parameter space than in a high-dimensional parameter space. For example, if a model with two parameters can fit the data equally well compared with a model with four parameters, the chance of the two-parameter model being true is larger than the four-parameter model. Such a penalty for having a higher dimensional parameter space in a complex model is reflected mathematically by the integration of the likelihood function over parameter space. Therefore, using the marginal likelihood value as a model selection criterion does not require a specific penalty term for model complexity.

However, for most models, it is analytically difficult to perform the integration to calculate the marginal likelihood $P(\mathbf{y}|M)$. The dimension of the parameter space can be high and the marginal likelihood can be difficult to express in a simple mathematical form. Therefore, the choice of the mathematical form for the posterior distribution is often limited or approximated to the normal distribution for computational convenience.

In recent years, iterative simulation methods have been developed to draw samples of the parameter values from general distributions [45]. These sampling methods are numerical techniques to obtain the posterior distribution of the parameters $p(\boldsymbol{\theta}|\mathbf{y}, M)$ in (2.6). When the posterior distribution is analytically difficult to calculate, the idea of these iterative *sampling methods* is to draw a set of samples $\boldsymbol{\theta}^{(i)}$ (where i represent i th sample of $\boldsymbol{\theta}$, $i = 1, 2, \dots, N$) independently from a sequence of distributions that converge, as iterations continue, to the desired target posterior distribution of $p(\boldsymbol{\theta}|\mathbf{y}, M)$, known as *Monte Carlo* integration. The reliability of the estimation from the Monte Carlo methods increases with the increased number of samples. The problem is that generating independent samples $\boldsymbol{\theta}^{(i)}$ can be difficult. When direct sampling is difficult, a *Markov chain* sequence of random samples can be drawn instead, which is defined by giving an initial distribution for $\boldsymbol{\theta}^{(0)}$, and the transition probability for $\boldsymbol{\theta}^{(i)}$ given the value for $\boldsymbol{\theta}^{(i-1)}$ [58]:

$$\boldsymbol{\theta}^{(i+1)} \sim p(\boldsymbol{\theta}^{(i+1)}|\boldsymbol{\theta}^{(i)}), i = 1, 2, \dots \quad (2.8)$$

The sample $\boldsymbol{\theta}^{(i+1)}$ only depends on the previous sample $\boldsymbol{\theta}^{(i)}$. Different *Markov Chain Monte Carlo* (MCMC) algorithms [44, 45] have a different way of determining whether a proposed new sample should be accepted. When a new sample $\boldsymbol{\theta}^{(i+1)}$ is accepted, the next proposed sample will be based on this new sample; when a new sample is rejected, the next proposed sample will be still based on the old sample. As $i \rightarrow \infty$, the Markov

chain converges to its target distribution $p(\boldsymbol{\theta}|\mathbf{y}, M)$. Introduced by Metropolis (1953) and widely popularised in the 1990s, MCMC remains one of the most important tools for Bayesian inference due to its flexibility in sampling from general distributions, and different MCMC algorithms have been developed to optimise the sampling procedures for different problems [45]. However, being iterative sampling methods, MCMC methods also have their shortcomings:

- 1) the samples from the exact solution are obtained at significant computational expense;
- 2) even though in theory the probability distribution will converge as i tends to infinity, in practice, it is not easy to decide how many samples are enough to be sure that the posterior distribution has converged;
- 3) standard MCMC algorithms such as the Metropolis-Hastings algorithm [44] and Gibbs algorithm [45] avoid calculating the marginal likelihood by choosing an acceptance ratio that can cancel out this term, and obtain the posterior distribution of the parameters without knowing the marginal likelihood. Since the model selection task requires comparison among the marginal likelihood values of competing models, standard MCMC algorithms cannot perform this task.

Therefore, standard MCMC algorithms such as Metropolis-Hastings algorithm and Gibbs algorithm are used for parameter estimation only. To obtain the marginal likelihood, more advanced MCMC algorithms (such as Population MCMC and thermodynamic integration [48, 59]) are required, which can be more computationally demanding.

Sequential Monte Carlo (SMC) methods [60] belong to another class of sampling methods that compute the posterior distribution of the parameters. Instead of constructing a Markov chain of samples from the posterior distribution like MCMC, SMC methods construct a sequence of distributions where the initial distribution is a simple distribution from which it is easy to sample and the final distribution is the posterior distribution. The intermediate distributions are required to be similar to each other, with the later distribution usually sampled from the previous distribution by using Important Sampling methods [61]. Importance sampling methods sample from a distribution that is different from the target distribution and compensate by weighting the samples so that the weighed samples would form a distribution that is closer to the target. Compared

to MCMC, SMC is better at handling multimodal distributions; however, as a sampling method, it has a similar problem to MCMC in terms of being computationally expensive. This is mainly because many intermediate distributions are needed to obtain the final distribution [62].

As opposed to the sampling methods, approximate (deterministic) methods, such as the expectation maximisation method [63] or expectation propagation method [64], do not need samples from the posterior probability distribution of the parameters. One of these methods is the variational Bayesian (VB) method, which has been proposed [65] to approximate the marginal likelihood. In the VB method, the marginal likelihood is often referred to as the *model evidence*. Being a deterministic method, the computational expense of the VB method is significantly less than the MCMC methods. MCMC methods take samples from the true posterior distribution of the parameters, and therefore with infinite sample sizes, the true posterior distribution can be obtained. The issue with MCMC methods is knowing a sufficient sample size in order that the sampled distribution is close enough to the true distribution. The VB method, on the other hand, approximates the true distribution using statistical distributions in the exponential family. Therefore, the issue with the VB approach is knowing the distance between the approximated posterior distribution and the true distribution. In this thesis, our main focus is to investigate applications of the VB method.

2.3 Variational Bayesian scheme

This section explains the key methodology involved in the VB algorithms, and modifications that were made to adapt the methodology to the applications of Chapters 3 and 4. As the model evidence is the key quantity that is used to compare different models in the task of model selection, this section emphasises how the VB method approximates the model evidence $P(y|M)$. Since approximated distributions are required in this method, a measure of the distance between two distributions, known as the *Kullback–Leibler divergence* (K-L divergence), is introduced first, and then to minimise this distance, the concept of free energy is introduced in the following sections.

2.3.1 Kullback–Leibler divergence and free energy

Biological systems adapt to the changing environment by constraining themselves via feedback mechanisms to remain at certain states that require a minimal level of energy [66]. To maintain a low energy consumption, a system should have a high probability of staying in states requiring low energy, and a low probability of staying in states requiring higher energy. In information theory, the amount of information each state contains is measured by a quantity termed ‘self-information’ [67]. Self-information, denoted by S , can be defined as the negative log-probability of an outcome state x :

$$S = -\log p(x) \quad (2.9)$$

If an outcome state rarely happens, e.g, a coin standing on its edge, the ‘self-information’ quantity will be large. A system’s average amount of information, defined as the *Shannon entropy* (referred to as ‘entropy’ for short in this thesis) [68], \mathcal{H} , can be defined as follows:

$$\mathcal{H} = -\langle \log p(x) \rangle_p = - \int p(x) \log p(x) dx \quad (2.10)$$

where $\langle \log p(x) \rangle_p$ represents the expectation of $\log p(x)$ with respect to the probability density function $p(x)$. A low entropy means that the system is easy to predict on average. To calculate the system’s entropy, a strict mathematical representation is required to describe the probability density $p(x)$ of the states x . $p(x)$ is usually complex (intractable) and cannot be captured by a closed-form statistical distribution. Therefore, a tractable distribution $q(x)$ is constructed to approximate $p(x)$. The information loss introduced by the approximation, denoted as $KL[P||Q]$, can be calculated as follows:

$$KL[P||Q] = \int p(x) \log \frac{p(x)}{q(x)} dx \quad (2.11)$$

$KL[P||Q]$, known as the Kullback-Leibler divergence (K-L divergence), represents the ‘distance’ between the probability distributions from $q(x)$ to $p(x)$. It is a directed measure, which means that $KL[P||Q]$ is not equal to $KL[Q||P]$. According to the Gibbs’ inequality [69], the K-L divergence is always non-negative. A smaller value in $KL[P||Q]$ indicates a better approximation of $p(x)$. To achieve the closest approximation, the K-L divergence needs to be minimised by determining the optimal configuration of the parameters of $q(x)$ using an inference method, and minimising $KL[P||Q]$ or $KL[Q||P]$

can result in different optimised parameters. $KL[P\|Q]$, where $q(x)$ is the denominator of the log within the integrand, allocates a higher weight to the region where $q(x)$ is near zero but $p(x)$ is not. Thus minimising $KL[P\|Q]$ leads to $q(x)$ covering as much of the region that $p(x)$ covers as possible. On the contrary, minimising $KL[Q\|P]$ leads to $q(x)$ avoiding regions where $p(x)$ is small [70]. Therefore, the parameter inference methods using $KL[Q\|P]$ (such as the VB method) tend to find a $q(x)$ that approximates a mode of $p(x)$; whereas the parameter inference methods using $KL[P\|Q]$ (such as the expectation propagation method [64]) tend to find a $q(x)$ that approximates the mean of $p(x)$. In the case of approximating the model evidence $P(\mathbf{y}|M)$, the distribution is often multi-modal due to high dimensional parameter space, with most of the posterior mass concentrated in several small regions of the parameter space, such as the two peaks of P as shown in Fig. 2.1. When such multi-modal distributions are approximated based on minimising $KL[Q\|P]$, one of these modes $q_1(x)$ will be found, depending on the prior distribution; on the other hand, when $KL[P\|Q]$ is minimised, the resulting approximation $q_2(x)$ tends to average across all the modes, but at the mean value, the probability density might be low. The mode of the parameter probability density is of interest in this thesis; the ‘closeness’ between two probability distributions is therefore defined as $KL[Q\|P]$. To avoid the problem of using $KL[Q\|P]$ – only one mode of the posterior distribution can be found even if the true posterior distribution is multi-modal – different values of the priors are set up to explore the parameter space.

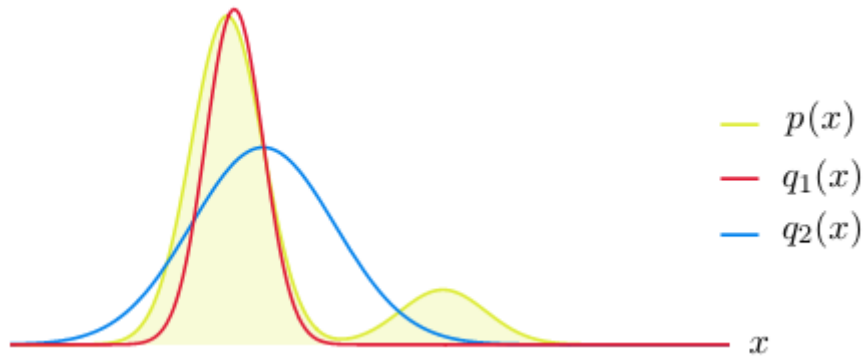


FIGURE 2.1: Approximation of the parameter distribution of P based on two K-L divergence measures; $q_1(x)$ is based on $KL[Q\|P]$ and $q_2(x)$ is based on $KL[P\|Q]$.

This method has been applied to approximate the probability distribution of system states \mathbf{x} in both applications considered in this thesis in the following manner. Assume \mathbf{x} can be observed through a series of measurements \mathbf{y} . Similar to (2.11), the

$KL[Q||P]$ divergence between the approximated distribution $q(\mathbf{x})$ and the true distribution $p(\mathbf{x}|\mathbf{y}, M)$, given observations \mathbf{y} and the model M , can be written as follows:

$$KL[Q||P] = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y}, M)} d\mathbf{x} = - \int q(\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{y}, M)}{q(\mathbf{x})} d\mathbf{x} \quad (2.12)$$

Then the conditional distribution $p(\mathbf{x}|\mathbf{y}, M)$ is replaced with a joint distribution $p(\mathbf{x}, \mathbf{y}|M)$ and the model evidence $p(\mathbf{y}|M)$ [41]:

$$p(\mathbf{x}|\mathbf{y}, M) = \frac{p(\mathbf{x}, \mathbf{y}|M)}{p(\mathbf{y}|M)} \quad (2.13)$$

So the $KL(Q||P)$ can be separated into two parts:

$$KL[Q||P] = \log p(\mathbf{y}|M) - \int q(\mathbf{x}) \log \frac{p(\mathbf{x}, \mathbf{y}|M)}{q(\mathbf{x})} d\mathbf{x} \quad (2.14)$$

The first term of the right side of (2.14) is the self-information of \mathbf{y} , which is also the logarithm of the model evidence and therefore known as the *log-evidence* of the model, representing the log-probability of obtaining the measurements given the model. When there is no prior preference over different models, the model with the highest log-evidence is the best model.

The second term of the right side of (2.14) is defined as the ‘*variational free energy*’ \mathcal{F} (free energy for short in this thesis):

$$\mathcal{F} = \int q(\mathbf{x}) \log \frac{p(\mathbf{x}, \mathbf{y}|M)}{q(\mathbf{x})} d\mathbf{x} \quad (2.15)$$

Since the K-L divergence is non-negative [70], the value of the free energy is always smaller than or equal to the log-evidence. As the log-evidence is not a function of $q(\mathbf{x})$, looking for a probability distribution $q(\mathbf{x})$ that minimises $KL[Q||P]$ is equivalent to looking for $q(\mathbf{x})$ that maximises the free energy. If the free energy is equal to the log-evidence, the K-L divergence reaches zero, meaning that there is no discrepancy between the true posterior distribution and the approximated distribution.

Note that the ‘free energy’ in the thesis is not the same concept as the free energy used in thermodynamics. In this thesis, the free energy value can be negative as it is the lower bound of the log-evidence. To reach as close as possible to the log-evidence, the value

of the free energy is maximised using an iterative algorithm explained in Section 2.3.2, and to understand the composition of the free energy better, an example is provided in Section 2.3.4.

2.3.2 Parameter and state estimation

In Bayesian approaches, all unknown quantities are treated as random variables, and represented by probability distributions. Therefore the parameters are treated as variables which do not change over time in the VB method. The maximisation of the free energy \mathcal{F} in (2.15) can be achieved by optimising the distributions of the parameters $\boldsymbol{\theta}$ and the system states \mathbf{x} . Therefore, the optimisation procedure depends on the assumptions regarding the parameters and the states, i.e. whether the parameters have a standard form of distribution (such as a normal distribution), are independent of each other or are interacting with each other, and whether the states are measurable. From (2.12), the posterior distributions of the states are approximated by $q(\mathbf{x})$. Assuming that $\boldsymbol{\theta}$ is the parameter vector for the system, the log-evidence in (2.14) can be approximated as follows:

$$\begin{aligned} \log p(\mathbf{y}|M) &\geq \int q(\mathbf{x}, \boldsymbol{\theta}) \log \frac{p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}|M)}{q(\mathbf{x}, \boldsymbol{\theta})} d\mathbf{x} d\boldsymbol{\theta} \\ &= \langle \log p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}|M) \rangle_{q(\mathbf{x}, \boldsymbol{\theta})} - \langle \log q(\mathbf{x}, \boldsymbol{\theta}) \rangle_{q(\mathbf{x}, \boldsymbol{\theta})} \\ &= \mathcal{E} + \mathcal{H} \\ &= \mathcal{F}(q(\mathbf{x}, \boldsymbol{\theta})) \end{aligned} \tag{2.16}$$

where $\langle \cdot \rangle_q$ denotes expectation with respect to the probability distribution q . The free energy \mathcal{F} comprises an energy term $\mathcal{E} = \langle \log p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}|M) \rangle_{q(\mathbf{x}, \boldsymbol{\theta})}$ and an entropy term $\mathcal{H} = -\langle \log q(\mathbf{x}, \boldsymbol{\theta}) \rangle_{q(\mathbf{x}, \boldsymbol{\theta})}$ as shown in (2.16). $q(\mathbf{x}, \boldsymbol{\theta})$ is an approximation of the distribution $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}, M)$. Driving $q(\mathbf{x}, \boldsymbol{\theta})$ closer towards $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}, M)$ by optimising the parameters of the distribution $q(\mathbf{x}, \boldsymbol{\theta})$ can push the value of \mathcal{F} towards the log-evidence $\log p(\mathbf{y}|M)$.

As introduced in Section 2.1.2, the precision of the system and measurement noise in (2.2), α_η and α_ε (inverses of the noise intensities), are the stochastic parameters of the model. Instead of treating the precisions of both types of noise as fixed values, they are modelled as Gamma-distributed variables with shape hyperparameters — a_η and a_ε and rate hyperparameters — b_η and b_ε . When the parameters of the model (α_η and

α_ε) are further parameterised, a_η , a_ε , b_η and b_ε are referred to as hyperparameters, and the model is referred to as a hierarchical model. Hierarchical models allow the prior distributions of certain parameters to be expressed as a probability distribution instead of a fixed value, and therefore one more hierarchy is added in the inference procedure of the noise in (2.2) to avoid giving priors to the noise precision that are too assertive. The prior values assigned to the hyperparameters are known as hyperpriors. Even though the hyperpriors are still specified as fixed values, the estimation of the parameters is less sensitive to a misspecification in a hyperprior than in a prior.

There are five components that need to be inferred from the data \mathbf{y} , and they can be denoted as a set of components (including the parameters and the states):

$$\Theta = \{\alpha_\eta, \alpha_\varepsilon, \boldsymbol{\theta}, \mathbf{x}_0, \mathbf{x}_{1:T}\} \quad (2.17)$$

α_η and α_ε are the stochastic parameters representing the precisions of system and measurement noise respectively; $\boldsymbol{\theta}$ is the deterministic parameter vector; \mathbf{x}_0 is the vector of initial states; $\mathbf{x}_{1:T}$ is the vector of the states: $\mathbf{x}_{1:T} = (x_{1:T}, \dot{x}_{1:T}, \dots, x_{1:T}^{(n-1)})^\top$ (the number in the parenthesis represents the order of the derivative) where each element in the vector represents the whole of the state trajectory. The main idea of the VB algorithm is to iteratively optimise the components of Θ : optimising one component while keeping the other four components fixed. To proceed to perform the iteration, the components are assumed to be conditionally independent for the given measurement data [41]. Therefore, individual component distributions $q(\Theta_i)$ can be used to approximate the combined distribution of all the components $q(\Theta)$, and this is known as the *mean-field approximation*. This treatment has been applied in the VB method to approximate the mixed posterior distribution of $q(\Theta)$:

$$q(\Theta) = \prod_i q(\Theta_i) = q(\alpha_\eta)q(\alpha_\varepsilon)q(\boldsymbol{\theta})q(\mathbf{x}_0)q(\mathbf{x}_{1:T}) \quad (2.18)$$

2.3.3 Iterative updating rule for hyperparameters, parameters and the states

Using the mean-field approximation, all components can be iteratively updated as illustrated in Fig. 2.2. At each iteration step, the probability distribution of one component, $q(\Theta_i)$ (q_i for short), is optimised to maximise the free energy \mathcal{F} , when the probability

distribution of the other components $q_{\setminus i}$ (the sign ‘\’ means ‘excluding’, i.e. all other probability distributions of Θ , except the i th component, are considered) are fixed. The cycle of the iterations continues until all the posterior distributions of the components are optimised and the free energy is maximised. Local convergence of the procedure to an optimal posterior distribution has been proved analytically [71].

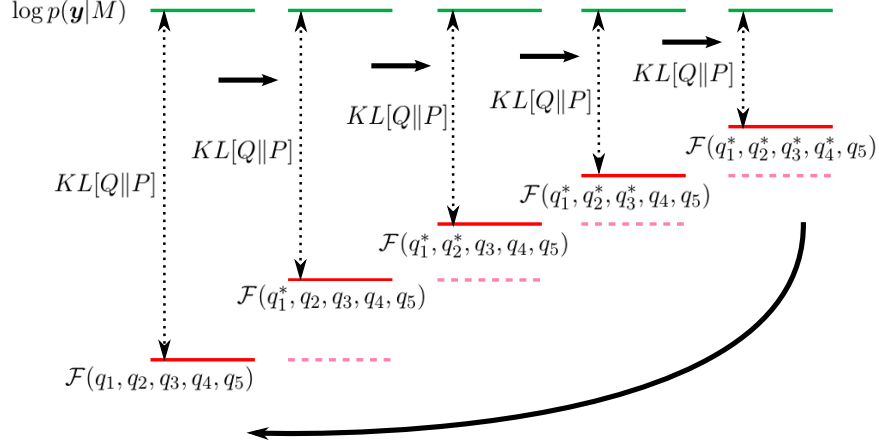


FIGURE 2.2: One iteration cycle of the VB algorithm. In each step, the probability distribution of the i th component q_i ($i = 1, 2, 3, 4, 5$) is optimised to maximise the free energy \mathcal{F} while keeping the probability distributions of the other components fixed. The lower bound of the log-evidence, i.e. the free energy, is guaranteed to increase (or stay unchanged) during each step. The mark * indicates that the probability distribution has been updated.

Computationally, to update the probability distribution of each component, variational calculus is applied. Variational calculus is a method to find a function (in this case $q(\Theta_i) = \{q(\alpha_\eta), q(\alpha_\varepsilon), q(\theta), q(\mathbf{x}_0), q(\mathbf{x}_{1:T})\}$, where each component approximates the probability densities of $p(\alpha_\eta|\mathbf{y}, M)$, $p(\alpha_\varepsilon|\mathbf{y}, M)$, $p(\theta|\mathbf{y}, M)$, $p(\mathbf{x}_0|\mathbf{y}, M)$, $p(\mathbf{x}_{1:T}|\mathbf{y}, M)$ respectively), that maximises a functional (in this case $\mathcal{F}(q(\alpha_\eta, \alpha_\varepsilon, \theta, \mathbf{x}_0, \mathbf{x}_{1:T}))$).

By equating the derivative of the free energy \mathcal{F} with respect to the probability distribution of one component $q(\Theta_i)$ to zero, $q(\Theta_i)$ can be obtained in the following form:

$$q(\Theta_i) = \arg \max_{q(\Theta_i)} \mathcal{F} \Rightarrow \frac{d\mathcal{F}}{dq(\Theta_i)} = 0 \quad (2.19)$$

where $q(\Theta_i)$, denoted as q_i for short, is the probability density function of the i th component of Θ , and $q_{\setminus i}$ (without i) are the probability density functions of the other four components.

Then $\log(q_i)$ can be obtained in the following form (its derivation can be found in [35]):

$$\begin{aligned} \log(q_i) &= \frac{1}{Z_i} \int \log p(\alpha_\eta, \alpha_\varepsilon, \boldsymbol{\theta}, \mathbf{x}_0, \mathbf{x}_{1:T}, \mathbf{y}|M) q_{\setminus i} \prod_{\setminus i} d\boldsymbol{\Theta}_{\setminus i} \\ &= \frac{1}{Z_i} \langle \log p(\alpha_\eta, \alpha_\varepsilon, \boldsymbol{\theta}, \mathbf{x}_0, \mathbf{x}_{1:T}, \mathbf{y}|M) \rangle_{q_{\setminus i}} \end{aligned} \quad (2.20)$$

where Z_i is a normalisation constant ensuring that q_i is a probability density function which integrates to one over the entire space. The VB algorithm iteratively optimises the components from $i = 1$ to 5 until \mathcal{F} is maximised. In (2.20), the terms inside the angle bracket represent the log-joint distribution of the measurements \mathbf{y} and all the components given in the model, $\log p(\alpha_\eta, \alpha_\varepsilon, \boldsymbol{\theta}, \mathbf{x}_0, \mathbf{x}_{1:T}, \mathbf{y}|M)$. In practice, however, not all the components are required to update q_i . Thus, (2.20) can be further simplified according to Fig. 2.3 in the following way. Take the update of $q(\alpha_\varepsilon)$ as an example. The measurements \mathbf{y} , defined as the ‘children’ of α_ε , are under direct influence. The measurements \mathbf{y} are also influenced by another ‘parent’ $\mathbf{x}_{1:T}$. The ‘child’ \mathbf{y} , and its ‘parent’ $\mathbf{x}_{1:T}$ are defined as the ‘Markov blanket’ of α_ε [72]. Therefore, the updating rule for $q(\alpha_\varepsilon)$ can be simplified by only considering the components that are under the Markov blanket of α_ε as follows:

$$\log(q_2) = \log(q(\alpha_\varepsilon)) = \frac{1}{Z_2} \langle \log p(\alpha_\varepsilon, \mathbf{x}_{1:T}, \mathbf{y}|M) \rangle_{q(\mathbf{x}_{1:T})} \quad (2.21)$$

The function $\langle \log p(\alpha_\varepsilon, \mathbf{x}_{1:T}, \mathbf{y}|M) \rangle_{q(\mathbf{x}_{1:T})}$ is defined as the ‘*variational energy*’ for α_ε , and is denoted as \mathcal{I}_{q_2} (the index ‘2’ refers to the second component of the vector $\boldsymbol{\Theta}$).

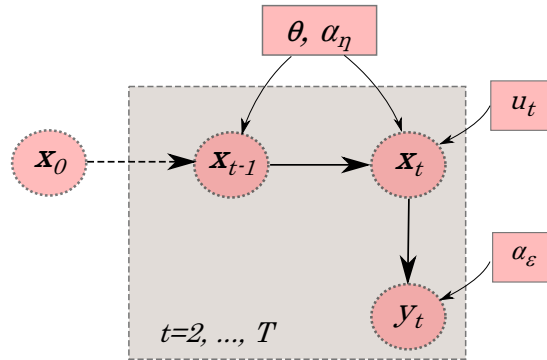


FIGURE 2.3: Illustration of the model M . y_t is the measurement point at time t , and \mathbf{x}_t is the state vector of the system at time t . $\boldsymbol{\theta}$ is the parameter vector of the system equation (2.2). α_η is the precision of the system noise and α_ε is the precision of the measurement noise, u_t is the input of the system at time t .

The VB iterative updating rule for the first two components (stochastic parameters α_η

and α_ε) is different from the other three components which are presented later in this section. As explained in Section 2.3.2, α_η and α_ε are further parameterised by the hyperparameters a_η , b_η , a_ε and b_ε . Therefore, α_η and α_ε are updated through updating the hyperparameters. Using the probability distribution of the parameters $\boldsymbol{\theta}$, the initial condition \mathbf{x}_0 and the states $\mathbf{x}_{1:T}$ estimated from the previous iteration cycle, the approximate posterior distribution probability of the noise precision can be obtained by using (2.20) directly, e.g. without further approximation of (2.20). Take the updating rule of $q(\alpha_\varepsilon)$ as an example. From (2.21), the joint probability function $p(\alpha_\varepsilon, \mathbf{x}_{1:T}, \mathbf{y}|M)$ can be further factorised into $p(\alpha_\varepsilon|M)$, $p(\mathbf{y}|\alpha_\varepsilon, \mathbf{x}_{1:T}, M)$, and $p(\mathbf{x}_{1:T}|M)$, out of which the first two terms are kept to optimise the variational energy of $q(\alpha_\varepsilon)$. The first term $p(\alpha_\varepsilon|M)$ is the prior of the measurement noise precision, which is modelled as a Gamma distribution. The second term, $p(\mathbf{y}|\alpha_\varepsilon, \mathbf{x}_{1:T}, M)$, is the conditional likelihood function (conditioned on the states $\mathbf{x}_{1:T}$), which is modelled as a Gaussian distribution. Integrating the conditional likelihood function $p(\mathbf{y}|\alpha_\varepsilon, \mathbf{x}_{1:T}, M)$ with respect to the normally distributed $q(\mathbf{x}_{1:T})$ gives the likelihood function for α_ε , $p(\mathbf{y}|\alpha_\varepsilon, M)$, which is normally distributed. The Gamma prior is conjugated to the Gaussian likelihood (‘conjugate’ indicates that the mathematical form of the posterior stays the same as the prior distribution after it has been updated by the likelihood function), resulting in a Gamma distributed posterior distribution $q(\alpha_\varepsilon)$.

For the other three components (the parameters $\boldsymbol{\theta}$, the initial state \mathbf{x}_0 and the states $\mathbf{x}_{1:T}$), the procedure is more complicated and requires further approximation of the variational energy. The variational energy for the parameter $\boldsymbol{\theta}$, denoted as \mathcal{I}_{q_3} , is shown below:

$$\log q(\boldsymbol{\theta}) = \frac{1}{Z_3} \mathcal{I}_{q_3} = \langle \log p(\alpha_\eta, \boldsymbol{\theta}, \mathbf{x}_0, \mathbf{x}_{1:T}, \mathbf{y}|M) \rangle_{q(\alpha_\eta), q(\mathbf{x}_0), q(\mathbf{x}_{1:T})} \quad (2.22)$$

Factorising the joint probability term $p(\alpha_\eta, \boldsymbol{\theta}, \mathbf{x}_0, \mathbf{x}_{1:T}, \mathbf{y}|M)$, three terms depend on $\boldsymbol{\theta}$ ($p(\boldsymbol{\theta}|M)$, $p(\mathbf{x}_{1:T}|\alpha_\eta, \boldsymbol{\theta}, \mathbf{x}_0, M)$, and $p(\mathbf{y}|\alpha_\eta, \boldsymbol{\theta}, \mathbf{x}_0, \mathbf{x}_{1:T}|M)$) and therefore are kept to optimise $q(\boldsymbol{\theta})$. The first item $p(\boldsymbol{\theta}|M)$ is modelled as a Gaussian distribution. Integrating the multiplication of the other two terms $p(\mathbf{x}_{1:T}|\alpha_\eta, \boldsymbol{\theta}, \mathbf{x}_0, M)p(\mathbf{y}|\alpha_\eta, \boldsymbol{\theta}, \mathbf{x}_0, \mathbf{x}_{1:T}|M)$ with respect to the probability distributions $q(\alpha_\eta)$, $q(\mathbf{x}_0)$, and $q(\mathbf{x}_{1:T})$ gives a Gamma distributed likelihood for $\boldsymbol{\theta}$. A Gaussian prior does not conjugate with a Gamma likelihood function, and therefore an approximation of the likelihood function $p(\mathbf{y}|\boldsymbol{\theta}, M)$ is needed to update $q(\boldsymbol{\theta})$. The updating rule for each of these three components (the parameters

θ , the initial state \mathbf{x}_0 and the states $\mathbf{x}_{1:T}$) needs such an approximation, known as the ‘VB-Laplace’ approximation.

A VB-Laplace approximation is the *Laplace approximation* in the VB context. The original Laplace approximation of the probability distribution $q(\Theta)$ is to match its first two moments (mean and variance) with a Gaussian distribution. The first two moments of Θ is a vector of means for each component of Θ and the covariance matrix. Assume the mode of the probability distribution of $q(\Theta)$ is at Θ^* . The Taylor-expansion of the logarithm of $q(\Theta)$ around the mode Θ^* gives the following:

$$\log q(\Theta) = \log q(\Theta^*) - \frac{1}{2}(\Theta - \Theta^*)^\top \mathcal{H}(\Theta - \Theta^*) + \dots \quad (2.23)$$

where \mathcal{H} is called the ‘Hessian matrix’ of $q(\Theta)$, and it is defined as the matrix of the second-order partial derivatives of $q(\Theta)$:

$$\mathcal{H} = \frac{\partial^2 \log q(\Theta)}{\partial \Theta^2} \quad (2.24)$$

The Laplace approximation works well when the actual distribution of $q(\Theta)$ is close to a normal distribution. However, when the mode is not near the majority of the probability mass, the Gaussian approximation around its mode is far from representative of the posterior distribution.

In the VB context, such problems can be avoided by iteratively tuning the mode and the covariance matrix of an individual component of Θ until an optimal distribution q_i is achieved. This iterative optimisation of each component is allowed due to the mean-field approximation. With the VB-Laplace approximation for each component (the parameters θ , the initial state \mathbf{x}_0 and the states $\mathbf{x}_{1:T}$, a numerical method, the Gauss-Newton method, can be applied to iteratively approach the optimised vector of means and the covariance matrix of q_i^* ($i = 3, 4, 5$) that maximise the respective variational energy (\mathcal{I}_{q_3} , \mathcal{I}_{q_4} and \mathcal{I}_{q_5}). A sequence of the vector of the mean values and the covariance matrix of q_i ($i = 3, 4, 5$) is obtained starting from the values from the last iteration cycle, and converging towards the optimised values q_i^* . Unlike Newton’s method, the Gauss-Newton method does not require the second derivative of the variational energy, i.e. the Hessian matrix of the variational energy is approximated. Therefore, it simplifies the computation compared with Newton’s method, but only works with mild nonlinearity in

the variational energy due to the approximation of the Hessian matrix. The maximised variational energies of the third component θ and the fourth component \mathbf{x}_0 can therefore be obtained to update the approximated posterior distribution functions of these two components. [72].

The VB-Laplace updating rule for the final component (the states $\mathbf{x}_{1:T}$) is the most complicated procedure among all the components. It is intuitive to calculate the approximate posterior probability of the states $\mathbf{x}_{1:T}$ simultaneously; however, the optimisation calculation involved exponentially increases with the number of states, which is undesirable when T is large. Therefore, instead of approximating the distribution of all the states simultaneously, a sequential propagation is used to evaluate $q(\mathbf{x}_t)$ point by point from $t = 1$ to $t = T$ using the extended Kalman-Rauch smoother algorithm [73]. This algorithm contains two passes that propagate the first and second order moments of the approximate posterior density of $q(\mathbf{x}_t)$. The first pass – the forward pass – is to compute the current state $q(\mathbf{x}_t)$ given the previous and the current observations $\mathbf{y}_{1:t}$. The second pass – the backward pass – is to incorporate the future observations into updating the posterior distribution on the current time step obtained by the forward pass. The details involved in these two steps are shown in [72] and in Chapter 5 of [41]. The VB-Laplace Kalman-Rauch algorithm applied to obtain the updating rule for the states is different from the traditional extended Kalman filter because it accounts for the uncertainties in the parameters Θ , which improves the performance when the VB-Laplace Kalman-Rauch algorithm deals with nonlinear systems with unknown parameters [72].

2.3.4 Free energy decomposition

As shown in Fig. 2.2, the iterative optimisation for the probability distributions of each component q_i ensures that the value of \mathcal{F} is monotonically increasing towards the model log-evidence until it converges, i.e. the difference between the free energy and the log-evidence – the K-L divergence – is minimised. The maximised free energy value \mathcal{F} can then be compared between different models to select the best model. The approximations of the posterior distributions of the hyperparameters, the parameters, the initial states and the states influence the value of the free energy \mathcal{F} , and therefore a closer look at the calculation of the free energy can provide more insights into how the estimation of each component influences the free energy value.

A simple example model is considered in this section to help understand the calculation of the free energy. Two unknown components, the parameter θ and the measurement noise precision α , are considered in this example:

$$\mathbf{z} = \theta \mathbf{t} + \varepsilon \quad (2.25)$$

where $\mathbf{z} = (z_1, z_2, \dots, z_N)^\top$, $\mathbf{t} = (t_1, t_2, \dots, t_N)^\top$, and ε corresponds to the measurement noise, which is modelled as a Gaussian distribution: $\varepsilon \sim \mathcal{N}(0, \alpha^{-1})$. The prior of the parameter θ is modelled as a Gaussian distribution: $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$, and the prior of the noise precision α is modelled as a Gamma distribution $\alpha \sim \mathcal{G}a(a_0, b_0)$, where a_0 and b_0 are the prior shape and rate hyperparameters for the noise precision. The posterior distribution of the parameter θ and the hyperparameters a and b are estimated through the iteration steps described in Section 2.3.3. At the n th iteration, the parameter θ and the hyperparameters (a and b) are denoted with the index n . According to (2.16), the free energy $\mathcal{F}(q(\alpha, \theta), \mathbf{z})$ can be decomposed as the energy term \mathcal{E} and the entropy term \mathcal{H} as follows:

$$\begin{aligned} \mathcal{F}(q(\alpha, \theta), \mathbf{y}) &= \mathcal{E} + \mathcal{H} \\ &= \langle \log p(\mathbf{z}, \theta, \alpha | M) \rangle_{q(\theta, \alpha)} - \langle \log q(\theta, \alpha) \rangle_{q(\theta, \alpha)} \\ &= \langle \log p(\alpha | M) \rangle_{q(\alpha)} + \langle \log p(\theta | M) \rangle_{q(\theta)} + \langle \log p(\mathbf{z} | \theta, \alpha, M) \rangle_{q(\theta, \alpha)} \\ &\quad - \langle \log q(\alpha) \rangle_{q(\alpha)} - \langle \log q(\theta) \rangle_{q(\theta)} \\ &= \underbrace{\langle \log \mathcal{G}a(\alpha | a_0, b_0) \rangle_{q(\alpha)}}_{(a)} + \underbrace{\langle \log \mathcal{N}(\theta | \mu_0, \sigma_0^2) \rangle_{q(\theta)}}_{(b)} + \underbrace{\langle \log \mathcal{N}(\mathbf{z} | \theta \mathbf{t}, \alpha^{-1}) \rangle_{q(\theta, \alpha)}}_{(c)} \\ &\quad - \underbrace{\langle \log \mathcal{G}a(\alpha | a_n, b_n) \rangle_{q(\alpha)}}_{(d)} - \underbrace{\langle \log \mathcal{N}(\theta | \mu_n, \sigma_n^2) \rangle_{q(\theta)}}_{(e)} \end{aligned} \quad (2.26)$$

The mean field approximation has been applied in (2.26) to factorise the energy term and the entropy term into three and two components respectively: (a), (b), and (c) comprise the energy term, and (d) and (e) constitute the entropy term. (2.26) can be

further expanded as follows:

$$\begin{aligned}
(a) : & \quad a_0 \log b_0 - \log \Gamma(a_0) + (a_0 - 1)(\psi(a_n) - \log b_n) - \frac{b_0 a_n}{b_n} \\
(b) : & \quad -\frac{1}{2} \log 2\pi - \log \sigma_0 - \frac{1}{2\sigma_0^2}(\mu_n - \mu_0)^2 \\
(c) : & \quad -\frac{N}{2} \log 2\pi + \frac{N}{2}(\psi(a_n) - \log b_n) - \frac{a_n}{2b_n} \mathbf{z}^T \mathbf{z} + \frac{a_n}{b_n} \mu_n \mathbf{t}^T \mathbf{z} - \frac{a_n}{2b_n}(\mu_n^2 + \sigma_n^2) \mathbf{t}^T \mathbf{t} \\
(d) : & \quad -\left(\log b_n - \log \Gamma(a_n) + (a_n - 1)\psi(a_n) - a_n \right) \\
(e) : & \quad -\left(\frac{1}{2} \log 2\pi - \log \sigma_n - \frac{1}{2} \right)
\end{aligned} \tag{2.27}$$

where $\Gamma(\cdot)$ is the gamma function, and $\psi(\cdot)$ is the digamma function, which is defined as $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$. The detailed derivation is as follows:

(a): The noise precision α follows a Gamma distribution with hyperparameters a_0 and b_0 : $\alpha \sim \mathcal{Ga}(a_0, b_0)$. The Gamma distribution can be expanded as $\alpha = \frac{b_0^{a_0}}{\Gamma(a_0)} \alpha^{a_0-1} e^{-b_0 \alpha}$. As the posterior mean of the noise precision $\alpha_{post} = a_n/b_n$, so $\langle \alpha \rangle_{q(\alpha)} = a_n/b_n$. Therefore, the log-probability of α with respect of $q(\alpha, \theta)$ can be obtained as:

$$\begin{aligned}
\langle \log \mathcal{Ga}(\alpha|a_0, b_0) \rangle_{q(\theta, \alpha)} &= a_0 \log b_0 - \log \Gamma(a_0) + (a_0 - 1) \langle \log \alpha \rangle_{q(\alpha)} - b_0 \langle \alpha \rangle_{q(\alpha)} \\
&= a_0 \log b_0 - \log \Gamma(a_0) + (a_0 - 1)(\psi(a_n) - \log b_n) - \frac{b_0 a_n}{b_n}
\end{aligned} \tag{2.28}$$

(b): The parameter θ follows a normal distribution with mean μ_0 and standard deviation σ_0 , and the updated mean of the parameter is μ_n , and therefore, the log-probability of θ with respect of $q(\alpha, \theta)$ can be obtained as:

$$\begin{aligned}
\langle \log \mathcal{N}(\theta|\mu_0, \sigma_0^2) \rangle_{q(\theta, \alpha)} &= -\frac{1}{2} \log 2\pi - \log \sigma_0 - \frac{1}{2\sigma_0^2}(\langle \theta \rangle_{q(\theta)} - \mu_0)^2 \\
&= -\frac{1}{2} \log 2\pi - \log \sigma_0 - \frac{1}{2\sigma_0^2}(\mu_n - \mu_0)^2
\end{aligned} \tag{2.29}$$

(c): The log-probability $\log p(\mathbf{y}|\theta, \alpha, M)$ with respect of $q(\alpha, \theta)$ can be obtained as:

$$\begin{aligned}
& \langle \log \mathcal{N}(\mathbf{z}|\theta\mathbf{t}, \alpha^{-1}) \rangle_{q(\theta, \alpha)} \\
= & -\frac{N}{2} \log 2\pi + \frac{N}{2} \langle \log \alpha \rangle_{q(\alpha)} - \frac{a_n}{2b_n} (\mathbf{z}^T \mathbf{z} - 2\langle \theta \rangle_{q(\theta)} \mathbf{t}^T \mathbf{z} + \langle \theta^2 \rangle_{q(\theta)} \mathbf{t}^T \mathbf{t}) \\
= & -\frac{N}{2} \log 2\pi + \frac{N}{2} (\psi(a_n) - \log b_n) - \frac{a_n}{2b_n} \mathbf{z}^T \mathbf{z} + \frac{a_n}{b_n} \mu_n \mathbf{t}^T \mathbf{z} - \frac{a_n}{2b_n} (\mu_n^2 + \sigma_n^2) \mathbf{t}^T \mathbf{t}
\end{aligned} \tag{2.30}$$

(d): The Shannon entropy of α is obtained as follows:

$$-\langle \log q(\alpha) \rangle_{q(\alpha)} = -\log b_n + \log \Gamma(a_n) + (1 - a_n) \psi(a_n) + a_n \tag{2.31}$$

(e): The Shannon entropy of θ is obtained as follows:

$$-\langle \log q(\theta) \rangle_{q(\theta)} = \frac{1}{2} \log 2\pi + \log \sigma_n + \frac{1}{2} \tag{2.32}$$

The free energy criterion seeks a balance between the accuracy and the complexity of the model [74], and can be expressed as:

$$\mathcal{F}(M) = \text{Accuracy}(M) - \text{Complexity}(M) \tag{2.33}$$

In this example, term (c) in (2.27) accounts for the accuracy of the model and the sum $-\sum((a) + (b) + (d) + (e))$ is the term penalising the model complexity. The analysis of the accuracy and complexity terms of the free energy is not straightforward because a change in one estimated parameter can have an impact on several terms.

For the accuracy term (c) in (2.27), the mean and the variance of the parameter prior distribution, μ and σ_0 , and the hyperpriors a_0 and b_0 of the noise precision do not influence the accuracy term. The accuracy term is only determined by the posterior distributions of the parameters and the posterior values of the hyperparameters. The accuracy term (c) equals

$$\frac{N}{2} \log \frac{\alpha_n}{2\pi} - \frac{\alpha_n}{2} (\mathbf{z} - \theta\mathbf{t})^\top (\mathbf{z} - \theta\mathbf{t}) \tag{2.34}$$

A higher residue $(\mathbf{z} - \theta\mathbf{t})^\top(\mathbf{z} - \theta\mathbf{t})$ from the model corresponds to a smaller value of the accuracy term (2.34). When the posterior noise precision $\alpha_n = a_n/b_n$ is less than 2π , the accuracy term is less than zero, which may lead to a negative free energy value.

The complexity term is more difficult to analyse. Here we focus on the influence of increasing the dimension of the parameter space on the complexity term, assuming that the hyperparameters are fixed. The complexity term can therefore be written as follows:

$$\text{Complexity}(M) = \frac{1}{2}\sigma_0^{-2}(\mu_n - \mu_0)^2 + \log \frac{\sigma_0}{\sigma_n} + C \quad (2.35)$$

where C includes all the terms that do not contain information about the parameters θ . With the increased number of parameters, the variances in (2.35) become vectors:

$$\text{Complexity}(M) = \frac{1}{2}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_0) + \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_n|} + C \quad (2.36)$$

where $\boldsymbol{\Sigma}_0$ is the prior of the covariance matrix of the parameter vector $\boldsymbol{\theta}$. The subindices ‘0’ and ‘n’ indicate the prior and the posterior of the covariance matrix respectively. The symbol $|\cdot|$ represents the determinant of the matrix. The increased complexity due to a higher dimension of the parameter space manifests mainly in the following three ways:

- 1) Assuming the prior variances for each parameter are the same, a higher dimension of the parameter space leads to more terms in the expression $\frac{1}{2}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_0)$, and, therefore, to a higher value of this expression.
- 2) The ratio of the determinants of the prior to posterior covariance matrices, $\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_n|}$, also known as the Occam factor [48], increases. The determinants of the covariance matrices correspond to the volume spanned by their eigenvectors in parameter space [75]. An increased number of parameters would enable the model to fit more diverse patterns of data, resulting in a more ‘flexible’ model [76]. On the other hand, to represent certain data, the parameters need to be specified in a narrower region for a model with more parameters. When the number of parameters increases, the relative posterior volume $|\boldsymbol{\Sigma}_n|$ with respect to the prior volume $|\boldsymbol{\Sigma}_0|$ in parameter space decreases. A narrower region in the posterior distribution of the parameters indicates a more brittle model, and this is the penalty for a more accurate fitting.

- 3) A higher correlation between the posterior distributions of the parameters (covariance matrix) increases complexity. Considering a model with two parameters, the determinant of the posterior covariance matrix for the parameters θ_1 and θ_2 is given as follows [75]:

$$|\Sigma_n| = (1 - r^2)\sigma_{\theta_1}^2\sigma_{\theta_2}^2 \quad (2.37)$$

where r is the posterior correlation between the parameters θ_1 and θ_2 ; σ_{θ_1} and σ_{θ_2} are the posterior standard deviations of the two parameters. A higher correlation between the two parameters indicates a smaller value of $|\Sigma_n|$, implying a larger complexity penalty term. When the correlation between the parameters is high, the parameters cannot be estimated accurately, causing higher posterior variances of both parameters (σ_{θ_1} and σ_{θ_2}), which offsets the higher complexity penalty caused by the higher correlation. In that situation, the additional parameter might not cause a decrease in the free energy value, and an over complicated model may be falsely selected [75]. Therefore, when the posterior variances of the parameters are high, the free energy criterion tends to choose an over complicated model; and when the posterior correlations between the parameters are high, the free energy criterion biases towards a simpler model.

In this simple example, the calculation of the free energy involves five interactive terms, each of which is influenced by the priors and posteriors of the parameter and/or the noise precision. In the applications shown in Chapters 3 and 4, the calculation of the free energy involves more terms and is more complicated. In this thesis, all of the calculation is dealt with by the variational Bayesian toolbox [77], the use of which is illustrated in Section 2.8.

Before the parameters are estimated, identifiability analysis (introduced in Section 2.6) is required to check if individual parameters can be uniquely determined given the model and model observation. After obtaining the estimated parameter distributions, parameter sensitivity analysis (introduced in Section 2.7) needs to be applied, especially when the posterior variances for the parameters are high, to check how sensitively the system would respond to a small change in the parameters. Also, because of the highly interactive nature among different terms of the free energy, effects caused by unsatisfactory estimation may not be reflected in the final value of the free energy when different terms offset each other. Therefore, other model selection criteria are also considered, together

with the free energy criterion, to choose the most appropriate model for the biomedical applications in later chapters.

2.4 The criteria for model selection

As introduced in Section 2.2, a good model selection criterion selects a model that describes the data adequately without overfitting it. The VB method uses the lower bound of the model evidence as in Fig. 2.2, i.e. the maximised value of the free energy, to select the best model candidates. Compared with other model selection criteria in the literature such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) [47, 78], the free energy criterion assigns a heavy penalisation term due to the integration over the whole parameter space, which gives a bias towards over-simplified models. As shown in the previous section, the calculation of the free energy is complicated and influenced by all of the priors and the posteriors of the parameters and the states. In cases where the free energy is sensitive to the change in priors, trusting the free energy value blindly can be problematic. Therefore, in practice, instead of treating the free energy criterion as a definite rule, other criteria, alongside the free energy, should be considered to gain quantitative information about how well the models fit the data, especially when the ‘goodness-of-fit’ for the measurements between the models is close [76]. When conflicts between the criteria occur, which is not unusual, it depends on the modeller to decide which criteria to use and which results to trust.

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are two commonly used model selection criteria, and both are simplified versions of the free energy criteria (or K-L divergence). Akaike introduced the K-L divergence (see Section 2.3.1) as a fundamental basis for model selection, and started advocating the AIC in the mid-70s. The AIC is based on the idea of minimising the K-L divergence between the true probability distribution $p(\mathbf{y})$ and the approximated probability distribution $q(\mathbf{y}|M)$ of the measurements \mathbf{y} as follows (refer to (2.11)):

$$KL[P||Q] = \langle \log p(\mathbf{y}) \rangle_{p(\mathbf{y})} - \langle \log q(\mathbf{y}|M) \rangle_{p(\mathbf{y})} \quad (2.38)$$

The model-dependent part of the K-L divergence $KL[P||Q]$ is the second term:

$$KL_r = \langle \log q(\mathbf{y}|M) \rangle_{p(\mathbf{y})} \quad (2.39)$$

which is known as the relative K-L divergence. Assuming the distribution $q(y|M)$ is parameterised by $\boldsymbol{\theta}$, the relative K-L divergence can be written as follows:

$$KL_r = \langle \log q(\mathbf{y}|\boldsymbol{\theta}, M) \rangle_{p(\mathbf{y})} \quad (2.40)$$

Unlike the free energy criterion, the AIC only considers the most probable value of the parameters $\boldsymbol{\theta}$. The problem is that the parameters are unknown and the expectation with respect to $p(\mathbf{y})$ in (2.40) cannot be evaluated since $p(\mathbf{y})$ is unknown. Using the same measurement data \mathbf{y} to estimate the parameters and then integrating $\log q(\mathbf{y}|\boldsymbol{\theta}, M)$ with respect to the probability distribution of \mathbf{y} would have the bias of using the same data \mathbf{y} twice. To avoid this problem, it is assumed that the most probable parameter $\hat{\boldsymbol{\theta}}$ can be obtained by applying the ML estimation from a fictitious data vector \mathbf{x} with the same length and the same probability distribution as \mathbf{y} but independent from \mathbf{y} . Then the following equation can be considered as the approximation for $\log q(\mathbf{y}|M)$:

$$\log q(\mathbf{y}|M) = \langle \log q(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{x}}, M) \rangle_{p(\mathbf{x})} \quad (2.41)$$

The fictitious data \mathbf{x} are created to estimate the parameters, and the dependence of the fictitious data \mathbf{x} can be eliminated via the expectation operation $\langle \cdot \rangle_{p(\mathbf{x})}$ in (2.41). A second-order Taylor expansion of $\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{x}})$ around $\hat{\boldsymbol{\theta}}_{\mathbf{y}}$, $\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{x}})$ can be obtained as follows:

$$\log q(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{x}}, M) = \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{y}}, M) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_{\mathbf{x}} - \hat{\boldsymbol{\theta}}_{\mathbf{y}})\mathcal{H}(\hat{\boldsymbol{\theta}}_{\mathbf{y}})(\hat{\boldsymbol{\theta}}_{\mathbf{x}} - \hat{\boldsymbol{\theta}}_{\mathbf{y}}) \quad (2.42)$$

where $\mathcal{H}(\hat{\boldsymbol{\theta}}_{\mathbf{y}})$ is the Hessian matrix evaluated at $\hat{\boldsymbol{\theta}}_{\mathbf{y}}$. $\hat{\boldsymbol{\theta}}_{\mathbf{x}}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{y}}$ are the ML estimates of the parameters for the data \mathbf{x} and \mathbf{y} respectively. Using the fact that \mathbf{x} and \mathbf{y} have the same distribution and have the same length, KL_r – which defines the AIC – can be obtained by inserting (2.42) into (2.39) as follows (a formal derivation can be found in [79]):

$$\begin{aligned} \text{AIC} = KL_r &= \left\langle \langle \log q(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{x}}, M) \rangle_{p(\mathbf{x})} \right\rangle_{p(\mathbf{y})} \\ &= \log(p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{y}}, M)) - k \end{aligned} \quad (2.43)$$

where k is the number of the parameters in the model.

A simple example will be presented to demonstrate how the AIC can be calculated. In this example, assuming the measurements y_t are conditionally independent of each other given the model and the parameter vector $\boldsymbol{\theta}$, the marginal likelihood $P(\mathbf{y}|\boldsymbol{\theta}, M)$ can be factorised into T separate terms:

$$P(\mathbf{y}|M, \boldsymbol{\theta}) = \prod_{t=1}^T P(y_t|M, \boldsymbol{\theta}) \quad (2.44)$$

Assuming that the measurement noise is Gaussian distributed with zero mean and variance σ_ε^2 , (2.44) can be further decomposed as follows:

$$P(\mathbf{y}|M, \boldsymbol{\theta}) = \prod_{t=1}^T \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}(y_t - \hat{y}_t)^2\right) \quad (2.45)$$

where \hat{y}_t is the estimation of the value of y_t based on the estimated parameters $\hat{\boldsymbol{\theta}}$. Taking the natural logarithm of both sides of (2.45) yields:

$$\log P(\mathbf{y}|M, \boldsymbol{\theta}) = -\frac{T}{2} \log(2\pi\sigma_\varepsilon^2) - \frac{1}{2} \sum_{t=1}^T \left(\frac{y_t - \hat{y}_t}{\sigma_\varepsilon} \right)^2 \quad (2.46)$$

The logarithm of the marginal likelihood, $\log P(\mathbf{y}|M, \boldsymbol{\theta})$, often called the *log-likelihood*, is used instead of the likelihood itself to simplify the calculation (using summation instead of multiplication). The maximum value of $P(\mathbf{y}|M, \boldsymbol{\theta})$ can be achieved when the parameter mean and variance satisfy the following conditions (the ML estimation):

$$\frac{\partial \log P(\mathbf{y}|M, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \quad (2.47)$$

and

$$\frac{\partial \log P(\mathbf{y}|M, \boldsymbol{\theta})}{\partial \sigma_\varepsilon} = 0 \Rightarrow \sigma_\varepsilon^2 = \frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2 \quad (2.48)$$

Substituting the estimated variance in (2.46), the log-likelihood function can be calculated as follows:

$$\log P(\mathbf{y}|M, \boldsymbol{\theta}) = -\frac{T}{2} \log \left(\frac{y_t - \hat{y}_t}{\sigma_\varepsilon} \right)^2 + C \quad (2.49)$$

where C is a constant independent of the model. Substituting the log-likelihood function in (2.43) by (2.49), the AIC can be calculated as follows:

$$\text{AIC} = -\frac{T}{2} \log \left(\frac{y_t - \hat{y}_t}{\sigma_\varepsilon} \right)^2 - k + C \quad (2.50)$$

where $\left(\frac{y_t - \hat{y}_t}{\sigma_\varepsilon} \right)^2$ is defined as the *residual sum of squares* (RSS).

The AIC has been reported to perform poorly for small numbers of data points, which motivated the inclusion of a correction term as follows [79]:

$$\text{AICc} = \text{AIC} - \frac{k(k+1)}{N-k-1} \quad (2.51)$$

where N is the number of data points and k is the dimension of the parameter space. The AICc, known as the ‘corrected AIC’, penalises parameters more than AIC does and the two criteria become approximately equal for $N > k^2$. Both the AIC and AICc have a tendency to select an overfitted model from the competing model candidates, which originates from overfitting the fictitious sample \mathbf{x} to estimate the parameter vector $\hat{\boldsymbol{\theta}}_{\mathbf{x}}$. However, the data generating mechanisms in practice are often more complicated than any proposed model candidates, especially when the data are sparse. In these cases, the AIC or the AICc may perform better compared with the free energy criterion due to the tendency of the AIC or the AICc to select an overfitted model. Both the AIC and AICc have been found useful in many applications reported in the literature [47].

The Bayesian Information Criterion (BIC) is an extension of the AIC. The expectation with respect to the probability distribution of $\hat{\boldsymbol{\theta}}_{\mathbf{x}}$, $\langle \log q(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{x}}, M) \rangle_{p(\hat{\boldsymbol{\theta}}_{\mathbf{x}})}$, rather than the probability distribution of \mathbf{x} , $\langle \log q(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{x}}, M) \rangle_{p(\mathbf{x})}$, is used to estimate the log-evidence $\log q(\mathbf{y}|M)$. Viewing the estimated parameter vector $p(\hat{\boldsymbol{\theta}}_{\mathbf{x}})$ as the prior for the parameter, model evidence $q(\mathbf{y}|M)$ can be obtained as follows:

$$q(\mathbf{y}|M) = \int q(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{x}}, M) q(\hat{\boldsymbol{\theta}}_{\mathbf{x}}|M) d\hat{\boldsymbol{\theta}}_{\mathbf{x}} \quad (2.52)$$

An approximation of the integral can be obtained under two assumptions: 1) the probability distribution of the parameter is flat around the estimated value $\hat{\boldsymbol{\theta}}_{\mathbf{x}}$; 2) the probability of the parameter is independent of the length of the data. Expanding $q(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{x}})$ around $\hat{\boldsymbol{\theta}}_{\mathbf{y}}$ using a second-order Taylor series, KL_r can be estimated as follows (the

formal derivation can be found in [79]):

$$KL_r \approx \log q(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathbf{y}}, M) + \underbrace{\log q(\hat{\boldsymbol{\theta}}_{\mathbf{y}}) + \frac{k}{2} \log 2\pi - \frac{k}{2} \log N}_{\text{does not scale with } N} \quad (2.53)$$

As the sample size N tends to infinity, the terms that do not scale with the number of data points N can be neglected, yielding the BIC:

$$\text{BIC} = \log q(\mathbf{y}|\hat{\boldsymbol{\theta}}, M) - \frac{k}{2} \log N \quad (2.54)$$

The BIC does not need the prior to quantify the model evidence $q(\mathbf{y}|M)$ because the prior term does not scale with the number of data points as stated by assumption 2) in the last paragraph. When the number of data points N is greater than 15, the complexity term of the BIC, $\frac{k}{2} \log N$, is larger than the complexity term of the AIC, k , and as a result the BIC has less probability of overfitting the data. Since it neglects the terms that do not scale with the number of data points, the complexity term is smaller than the complexity term in the free energy criterion. Therefore, the BIC is less strict towards model complexity compared with the free energy criterion, but stricter compared with the AIC.

Depending on the chosen criterion (AIC or BIC), a higher value of the AIC or the BIC indicates a better model. The difference in either the AIC or the BIC between two models is assessed using the Bayes factor [80]. Advocated by Jeffreys in 1961, the Bayes factor aims at providing a Bayesian equivalent to hypothesis testing in classical statistics. When two models are compared, the probability of one model M_1 being true over the probability of another model M_2 being true can be obtained as follows:

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{p(M_1)}{p(M_2)} \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} \quad (2.55)$$

The prior probabilities of the two models are considered equal, i.e. $p(M_1) = p(M_2)$, and therefore the Bayes factor in favour of M_1 over M_2 is defined as follows:

$$B_{1,2} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} \quad (2.56)$$

Kass and Raftery [80] suggested interpreting the Bayes factor in the natural logarithm scale as in Table 2.1:

TABLE 2.1: Bayes factor compared between models M_1 and M_2

$\log(B_{1,2})$	$B_{1,2}$	Evidence against M_2
$0 \sim 2$	$1 \sim 3$	Not significant
$2 \sim 6$	$3 \sim 20$	Positive
$6 \sim 10$	$20 \sim 150$	Strong
> 10	> 150	Very strong

Bayes' rule works for the AIC, BIC, and the free energy criterion: the differences in the AIC values, the BIC values and the free energy values are equivalent to a log-Bayes factor $\log(B_{1,2})$ in Table 2.1. If the AIC is chosen to be the model selection criterion, then $\text{AIC}(M_1) - \text{AIC}(M_2) > 3$ (3 is a conventional choice based on [81] and [80]) indicates that M_1 outperforms M_2 . If the BIC is chosen to be the model selection criterion, then $\text{BIC}(M_1) - \text{BIC}(M_2) > 3$ indicates that M_1 outperforms M_2 . If the free energy criterion is chosen, then $\mathcal{F}(M_1) - \mathcal{F}(M_2) > 3$ indicates that M_1 outperforms M_2 .

Compared with the free energy criterion, both the AIC and BIC are more widely used due to their simplicity. They do not take the uncertainty about parameters into consideration, i.e. the value of the AIC or the BIC is only based on the estimated parameter values from the ML method. The uncertainty of the parameters is underestimated in both methods, and therefore overfitted models are more likely to be chosen using either of these two criteria [80], which has been shown by [82] and [83]. The free energy criterion, on the other hand, provides a model comparison taking the uncertainty of the parameters into account. It is worth noting that a higher free energy value does not guarantee higher model evidence because the gap between the free energy and the model evidence can be different for each model candidate. As shown in Chapter 4 in [41], the gap – the K-L divergence – is found to be positively correlated with the number of parameters, which means that the estimation of the model evidence becomes more pessimistic as the model complexity increases, rendering the VB methods to suffer from a tendency to select a model that underfits the data. Therefore, in later applications in Chapter 3 and Chapter 4, several criteria are considered together to select the best model, with the purpose of minimising the limitations of each criterion.

2.5 Priors and hyperpriors

As stated in Section 2.3.2, the VB method treats all unknown quantities as random variables, and if certain prior knowledge or information about these random variables (including hyperparameters, the parameters, the initial state or the states of the system) is known before accounting for the data, it should be taken into consideration. Mathematically, this information is encapsulated as the prior, $p(\theta|M)$, that is one factor of the numerator in (2.3) shown in Section 2.2.1. The essence of Bayesian inference is to incorporate the prior information into the information extracted from the data [84], and obtain the posterior combining both. The prior provides the opportunity for the modeller to express belief regarding the variable of interest, and should be treated with caution.

There are two categories of priors: informative priors and uninformative priors. Informative priors are usually chosen based on known information about the variable of interest. An inference without priors could lead to unreliable results, especially when the available data set is small. For example, if a coin was tossed three times and each time landed heads, the data by themselves would suggest a zero probability of obtaining tails, which contradicts common sense. Including an informative prior in the form of a normal distribution with a mean value of 0.5 would provide a more realistic predictive outcome of the experiment. Another type of informative prior can be originated from previously available data. If a model is updated when more data become available, then the posterior distribution of the parameters inferred from the previous data may be used as the prior distribution of the parameters for the data obtained later.

The second category is uninformative or weakly informative priors. Common weak priors include flat priors, priors with large variances, Jeffreys priors, etc [85]. A flat prior, i.e. the unknown parameter is uniformly-distributed from negative infinity to positive infinity, is often chosen to be the first attempt at an uninformative prior. However, such a prior is improper (does not integrate to one), and might lead to an improper posterior which invalidates the model [86]. An improper prior would also be problematic for calculating the model evidence since the model evidence is the conditional likelihood, $P(\mathbf{y}|\boldsymbol{\theta}, M)$, integrated over the prior distribution $p(\boldsymbol{\theta}|M)$ as shown in (2.7). Therefore, weakly informative priors, or ‘vague’ priors, rather than uninformative priors, are preferred in practice, and in this thesis, weakly informative priors, such as priors with large

variances as well as Jefferys priors, are chosen for all of the variables of interest Θ as in (2.17).

One choice of a weak prior is to assign a normal distribution with a large variance, such as $\theta \sim \mathcal{N}(0, 10^4)$ [84]. In both clinical applications in Chapter 3 and Chapter 4, the priors for the parameter vectors θ , the states $\mathbf{x}_{1:T}$ and the initial states \mathbf{x}_0 are set to be normally distributed with zero means and large variances of order 10^4 (larger variances in the priors have been tried to see if it makes a difference in the posterior distributions, and 10^4 is regarded as large for the two applications mentioned).

Another choice of a weak prior is a Jeffreys prior. The idea of a Jeffreys prior originated from the attempt to solve one of the problems of the uniform prior — a uniform distribution for a random variable does not imply a uniform distribution for a function of the random variable, i.e. an uninformative prior for the random variable becomes informative after re-parameterisation. For example, assume a uniform prior distribution for the parameter ρ which describes the probability of raining tomorrow. The probability of raining tomorrow and the day after tomorrow is $\vartheta = \rho^2$, and therefore $\vartheta \sim \frac{1}{2\sqrt{\vartheta}}$, which is not uniformly distributed any more. A Jeffreys prior was proposed so that the prior distribution can be invariant to such re-parameterisation [87]. It is based on the principle of maximising the K-L divergence between the prior and the posterior distribution [88]. A Jeffreys prior of parameter ρ is defined as follows:

$$p(\rho) \sim \mathcal{FI}(\rho)^{1/2} \quad (2.57)$$

where \mathcal{FI} is the ‘Fisher information matrix’ [89], defined as the expectation of the second derivative of the log-likelihood function with respect to the probability distribution of ρ , or the second moment of the log-likelihood function:

$$\mathcal{FI}(\rho) = -\left\langle \frac{d^2 \log p(\mathbf{y}|\rho, M)}{d\rho^2} \right\rangle_{p(\rho)} \quad (2.58)$$

The Fisher information matrix measures the curvature of the log-likelihood function. Since high curvature represents large changes in the likelihood when the value of the parameter changes, a Jeffreys prior gives more weight to those parameter values ensuring that maximised information can be obtained from the influence of the data [84, 90]. Applying (2.58) to obtain a Jeffreys prior for the precision parameter α in the example

shown in (2.26):

$$p(\alpha) \sim \alpha^{-1} \quad (2.59)$$

Given (2.59), it can be easily proved that the natural logarithm of α is uniformly distributed. Applying a change of variable $\beta \log \alpha$ yields:

$$\int p(\beta) d\beta = \int p(\alpha) \frac{d\alpha}{d\beta} d\beta = \int \frac{1}{C} e^{-\beta} e^{\beta} d\beta = 1 \quad (2.60)$$

where C is the normalisation constant. As β follows a uniform distribution implying the prior does not favour any one scale over another, after Jeffreys prior for α is therefore weakly informative.

In this thesis, the priors for the noise precision parameters, α_η and α_ε shown in Section 2.1.2, have been set to approximate Jeffreys priors by choosing the hyperparameters that define the Gamma distributed noise precisions. Using the hyperparameters to specify the noise provides a more objective prior by estimating the noise precision as a probability distribution instead of a number. Even though the hyperpriors of the hyperparameters, a_0 and b_0 , are required to describe the distribution of the noise precision, it is less arbitrary than specifying a number as the prior of the noise precision using subjective information [85]. Consider measurement noise as an example. The measurement noise is a vector $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N)^T$, $t = 1, 2, \dots, N$, and ε_t at time t is the difference between the measured values y_t and the estimated values \hat{y}_t : $\varepsilon_t = y_t - \hat{y}_t$. The value of the noise precision α is drawn from the Gamma distribution $\mathcal{Ga}(a, b)$ and then the measurement noise ε_t is drawn from the Gaussian distribution $\mathcal{N}(0, \alpha^{-1})$. The joint probability $p(\varepsilon, \alpha | M)$ can be decomposed as $p(\varepsilon, \alpha | M) \propto p(\varepsilon | \alpha, M) p(\alpha | M)$, and further expanded as the multiplication of Gaussian distributions and a Gamma distribution:

$$p(\varepsilon, \alpha | M) = \prod_{t=1}^N \frac{\sqrt{\alpha}}{\sqrt{2\pi}} \exp\left(-\frac{\varepsilon_t^2 \alpha}{2}\right) \frac{b_0^{a_0}}{\Gamma(a_0)} \exp(-b_0 \alpha) \alpha^{(a_0-1)} \quad (2.61)$$

Grouping similar terms together, (2.61) can be written as:

$$p(\varepsilon, \alpha | M) \propto \alpha^{a_0 + \frac{N}{2} - 1} \exp\left(-\alpha \left(b_0 + \frac{1}{2} \sum_{i=1}^N \varepsilon_i^2\right)\right). \quad (2.62)$$

which is in the form of a Gamma distribution, so $p(\boldsymbol{\varepsilon}|M)$ can be obtained by integrating α out in (2.62) as follows:

$$p(\boldsymbol{\varepsilon}|M) \propto \left(b_0 + \frac{1}{2} \sum_{i=1}^N \varepsilon_i^2 \right)^{-(a_0 + \frac{N}{2})}. \quad (2.63)$$

Dividing the joint distribution $p(\boldsymbol{\varepsilon}, \alpha|M)$ of (2.62) by the distribution for the measurement noise $p(\boldsymbol{\varepsilon})$, the posterior conditional distribution of $p(\alpha|\boldsymbol{\varepsilon}, M)$ can be obtained as follows:

$$p(\alpha|\boldsymbol{\varepsilon}, M) \propto \alpha^{a_0 + \frac{N}{2} - 1} \left(b_0 + \frac{1}{2} \sum_{i=1}^N \varepsilon_i^2 \right)^{a_0 + \frac{N}{2}} \exp \left(-\alpha \left(b_0 + \frac{1}{2} \sum_{i=1}^N \varepsilon_i^2 \right) \right). \quad (2.64)$$

As seen from (2.64), the posterior conditional distribution $p(\alpha|\boldsymbol{\varepsilon}, M)$ is in the form of a Gamma distribution with its shape and rate parameters $(a_0 + \frac{N}{2}, b_0 + \frac{1}{2} \sum_{i=1}^N \varepsilon_i^2)$. This means that the posterior conditional distribution depends on the hyperpriors a_0 and b_0 , the sample size N and the squared sum of the differences between the observations and the estimations. The prior mean of α is a_0/b_0 and the prior variance is a_0/b_0^2 . Therefore, if the values of a_0 and b_0 are small compared with the sample size N , the prior carries little information about α , and the posterior hyperparameters are approximately $(\frac{N}{2}, \frac{1}{2} \sum_{i=1}^N \varepsilon_i^2)$. The posterior mean of the noise precision is proportional to the inverse of α , i.e. $\frac{N}{2} / \frac{1}{2} \sum_{i=1}^N \varepsilon_i^2 = \alpha^{-1}$, which coincides with the Jeffreys prior as shown in (2.59).

To conclude, when the hyperpriors (a_0 and b_0) of the Gamma distribution are negligible compared to the number of the data points N , the Gamma distribution for the noise precision is approximately a Jeffreys prior as in (2.59). Therefore, $\mathcal{G}a(0.001, 0.001)$ or $\mathcal{G}a(0.01, 0.01)$ or $\mathcal{G}a(1, 1)$ are the common choices for the prior distribution of the noise precision [85]. However, when a_0 or b_0 are set too close to zero, it can lead to an improper posterior density: as $\mathcal{G}a(0, 0)$ is an improper prior and leads to an improper posterior, an approximation to the improper prior would also approximate its improper posterior, leading to an unstable inference [91]. As illustrated by [85], when the inferred variance α^{-1} is close to zero, the inference result becomes sensitive to a_0 and b_0 in (2.64). For the applications in Chapter 3 and Chapter 4, different hyperpriors, such as $\mathcal{G}a(0.001, 0.001)$, $\mathcal{G}a(0.01, 0.01)$, and $\mathcal{G}a(1, 1)$, have been used and compared.

There are different techniques to implement the hyperparameters' inference. In many empirical Bayesian methods, the hyperparameters are optimised before a regular Bayesian

inference is conducted with the estimated hyperparameters to infer the states and the parameters of the model. The drawback of these methods is that the same data are used twice, leading to data overfitting. The VB method includes the optimisation of the hyperparameters in the iterative algorithms and therefore successfully avoids this major drawback of empirical Bayesian methods.

Regarding the functional form of the priors, the VB method works better when the priors and the posteriors belong to the same probability family. If the posterior distribution is in the same family as the prior distribution, the prior is called a *conjugate* prior for the likelihood function. Because the posterior distribution is a standard statistical distribution, a conjugate pair of the prior and the likelihood distribution enables us to compute the posterior density analytically without the trouble of the integration operation to calculate the normalisation constant in the denominator (as shown in (2.6)). In this thesis, two conjugate pairs have been used: a Gaussian distributed prior and a Gaussian-distributed likelihood function are a conjugate pair; a Gamma-distributed prior and a Gaussian-distributed likelihood function are also a conjugate pair. Using conjugate pairs makes the mathematical calculation much easier, but at the same time limits the applications of the model. From the previous example, the prior of the noise precision $p(\alpha|M)$ and the posterior $p(\alpha|\varepsilon, M)$ are both in the form of a Gamma distribution, and the likelihood function $p(\varepsilon|\alpha, M)$ is in the form of a Gaussian distribution. Considering many likelihood functions are Gaussian distributed or at least belong to the exponential family, which is the case in the considered applications in this thesis, the requirement for the conjugation prior can be satisfied by many models, making the VB method a good inference choice for those cases. When the distributions of the parameters do not conjugate with the likelihood function, approximation of the distributions is required, which would introduce bias towards the inference results.

2.6 Structural identifiability

In biomedical systems modelling, perturbations through some forms of inputs, for example an injection or infusion of drug, are often applied to the physiological system of interest; measurement data can then be obtained as the output of the system. The system structure is an unknown black box and its parameters are learned through this input-output relationship. Before estimating the system parameters, it is necessary to

make sure that the parameters can be uniquely identified with respect to a particular input-output structure, and this is referred to as structural identifiability analysis [92].

Theoretical structural identifiability tests are based on two assumptions: 1) the structure of the model is appropriate; 2) there are no measurement errors. For the deterministic form of the model in (2.2), the generic parameter vector $\boldsymbol{\theta}$ is defined to be *structurally locally identifiable* if there exists a neighbourhood of vectors around $\boldsymbol{\theta}$, $\mathcal{N}(\boldsymbol{\theta})$, such that if $\bar{\boldsymbol{\theta}} \in \mathcal{N}(\boldsymbol{\theta})$ and for every input \mathbf{u} and $t \geq 0$, $\mathbf{y}(t, \boldsymbol{\theta}) = \mathbf{y}(t, \bar{\boldsymbol{\theta}})$, then $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$. If $\mathcal{N}(\boldsymbol{\theta})$ is the whole parameter space, the previous statement still holds true, then $\boldsymbol{\theta}$ is *structural globally identifiable*.

A number of techniques are available for performing the structural identifiability analysis of a linear model, but only the most common method, the Laplace transform approach, is applied here. The following is a summarisation from a review article [93]. Consider a general n -dimensional linear model in state-space form given by

$$\begin{aligned}\dot{\mathbf{x}}_t &= \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t\end{aligned}\tag{2.65}$$

where \mathbf{x} is the $n \times 1$ state vector, \mathbf{y} is the $m \times 1$ output vector and \mathbf{u} is the $r \times 1$ input vector; \mathbf{A} is the $n \times n$ state matrix, \mathbf{B} is the $n \times r$ input matrix, \mathbf{C} is the $m \times n$ output matrix. The initial conditions are assumed to be zero here, $\mathbf{x}_0 = 0$ (non-zero initial conditions will be considered in Chapter 4). Taking the Laplace transforms of the system (2.65) gives:

$$\begin{aligned}s\mathbf{X}(s) &= \mathbf{A}\mathbf{X}(s) + \mathbf{B}\mathbf{U}(s) \\ \mathbf{Y}(s) &= \mathbf{C}\mathbf{X}(s)\end{aligned}\tag{2.66}$$

where $\mathbf{X}(s)$, $\mathbf{U}(s)$ and $\mathbf{Y}(s)$ are the Laplace transform of the state, input and output vectors, respectively. Rearranging and combining these equations gives:

$$\mathbf{Y}(s) = \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}(s)\tag{2.67}$$

where $\mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}$ is the transfer function matrix. Measurements taken for the transfer function are assumed known so that the coefficients of the powers of s in the

numerators and denominators of the measured outputs can be assumed to be uniquely determined by the input-output relationship.

For example, for a two-state linear system:

$$\ddot{x}_t + \theta_1 \dot{x}_t + \theta_2 x_t = F u_t \quad (2.68)$$

with observation

$$y_t = x_t \quad (2.69)$$

and initial conditions

$$\begin{aligned} x_0 &= 0 \\ \dot{x}_0 &= 0 \end{aligned} \quad (2.70)$$

the Laplace transform of (2.68) is given by

$$(s^2 + \theta_1 s + \theta_2)X(s) = F U(s) \quad (2.71)$$

The Laplace transform of the observations is of the form:

$$Y(s) = \frac{F}{s^2 + \theta_1 s + \theta_2} U(s) \quad (2.72)$$

where F , θ_1 and θ_2 are assumed to be known [93]. Therefore, θ_1 , θ_2 and F are uniquely identifiable.

The identifiability analysis for a nonlinear system is generally more complicated than for linear systems, especially when the number of the parameters is large. By locally linearising nonlinear systems, local identifiability can be evaluated since this only requires the parameter to be unique in small neighbourhoods of the parameter space [94]. One of the widely accepted methods to check the identifiability of a nonlinear system is the *Taylor series approach* [93, 95, 96]. The idea is to evaluate the measurements y and their successive time derivatives $y^{(i)}$ at a particular time, usually $t = 0$, using a Taylor series. For the deterministic system considered in (2.1a) and (2.1b) with exact measurements, i.e. $y_t = x_t$, the observed measurements y can be expanded as a Taylor series around $t = 0$ as follows:

$$y_t = y_0 + t\dot{y}_0 + \frac{t^2}{2!}\ddot{y}_0 + \cdots + \frac{t^i}{i!}y_0^{(i)} + \cdots \quad (2.73)$$

The coefficients of the Taylor series $y_0^{(i)}$ are theoretically measurable and $y_0^{(i)}$ is a function of the system parameters θ . Therefore, the identifiability problem reduces to determining the possible solutions of the parameter vector θ that generates the infinite list of coefficients $y_0^{(i)}$ ($i = 0, 1, 2, \dots$). Margaria et al. [97] provides an upper bound on the number of coefficients that need to be considered with all polynomial transfer coefficients, which is $n + r$, where r is the number of parameters. If only one solution of the parameter vector exists, the parameters are uniquely identifiable. If the number of solutions is countable, the parameters are locally identifiable. If the number of solutions is uncountable, the parameters are unidentifiable [96]. To illustrate this method, a two-state deterministic system is considered with the states at time zero x_0 impulsively perturbed. The impulsive input can be expressed as initial conditions at $t = 0$ and the system in (2.1a) can be written as follows:

$$\ddot{x}_t = -(\theta_{k_2}x_t^2 + \theta_{k_1}x_t + \theta_{k_0})\dot{x}_t - \theta_1x_t \quad (2.74)$$

with known initial conditions:

$$x_0 = 0 \quad (2.75a)$$

$$\dot{x}_0 = F \quad (2.75b)$$

Substituting the initial conditions into (2.74),

$$\ddot{x}_0 = -(\theta_{k_2}x_0^2 + \theta_{k_1}x_0 + \theta_{k_0})\dot{x}_0 - \theta_1x_0 = -\theta_{k_0}F \quad (2.76)$$

the third derivative of the state x at time 0 can be obtained by differentiating (2.74),

$$\begin{aligned} x_0^{(3)} &= -(2\theta_{k_2}x_0\dot{x}_0 + \theta_{k_1}\dot{x}_0)\dot{x}_0 - (\theta_{k_2}x_0^2 + \theta_{k_1}x_0 + \theta_{k_0})\ddot{x}_0 - \theta_1\dot{x}_0 \\ &= -F(\theta_1 - \theta_{k_0}^2 + F\theta_{k_1}) \end{aligned} \quad (2.77)$$

Higher derivatives at $t = 0$ can be obtained by continually differentiating (2.77). Assuming another parameter vector

$$\bar{\theta} = \begin{pmatrix} \bar{\theta}_{k_0} \\ \bar{\theta}_{k_1} \\ \bar{\theta}_{k_2} \\ \bar{\theta}_1 \end{pmatrix} \quad (2.78)$$

can generate the same output (and the same derivatives of any order of the output) as θ . By comparing the derivative terms calculated using the parameter vector $\bar{\theta}$ and θ , (2.76) yields:

$$-\theta_{k_0} F = -\bar{\theta}_{k_0} F \quad (2.79)$$

Since F is the known initial condition which is non-zero, (2.79) implies that $\theta_{k_0} = \bar{\theta}_{k_0}$. Equation (2.77) yields:

$$-F(\theta_1 - \theta_{k_0}^2 + F\theta_{k_1}) = -F(\bar{\theta}_1 - \bar{\theta}_{k_0}^2 + F\bar{\theta}_{k_1}) \quad (2.80)$$

Three more equations can be obtained by equating higher order derivatives of x at $t = 0$ (including $x_0^{(4)}$, $x_0^{(5)}$, and $x_0^{(6)}$, here $n + r = 6$, and hence these terms are needed) using the parameter vectors $\bar{\theta}$ and θ in the same manner. These three equations, together with (2.79) and (2.80), can be solved through symbolic calculation by the Mathematica software [98], and the following unique solution is obtained:

$$\theta_{k_0} = \bar{\theta}_{k_0} \quad \theta_{k_1} = \bar{\theta}_{k_1} \quad \theta_{k_2} = \bar{\theta}_{k_2} \quad \theta_1 = \bar{\theta}_1 \quad (2.81)$$

Therefore, the illustrative nonlinear model is proved to be uniquely identifiable.

These aforementioned structural identifiability analyses provide a theoretical basis for deterministic linear system as well as nonlinear system and can be performed without any actual experimental observations. In biomedical applications, the uncertainty in the model and the measurement noise are often large. This means that even if the parameters are theoretically identifiable, fitting a model to data may still not yield unique and optimal parameter estimates, especially when the measurements are sparse and do not have enough constraining power over the parameters [99]. Unidentifiability causes no real difficulties [100, 101] for Bayesian approaches by assigning different priors in the parameter space to check for the uniqueness of the parameter estimates. When the parameters are not uniquely identifiable, the mapping from the distribution of the parameters to the marginalised likelihood function is not one-to-one, and the model evidence is multi-modal. In these situations, it is necessary to select different priors in the parameter space and check if these different priors yield the same posteriors.

The noise in stochastic nonlinear models plays an important role in identifying the deterministic parameters in the inference process. Meaningful parameter estimation can

only be obtained with a sufficiently small amount of noise in the measurements [102]. The noise intensity can be used as an indicator of how well the parameters describe the data. In this thesis, the main purpose is to build data-driven models to describe clinical data; taking account of the noise makes the model more realistic, and the description of the data more accurate. However, the main interest in both applications remains at the deterministic level which reveals the relationship between the inputs and outputs of the system, and may help predict clinical outcomes.

The theoretical parameter identifiability analysis is performed for the nonlinear deterministic system in Chapter 3 and the linear deterministic systems in Chapter 3 and Chapter 4. Different prior settings for the hyperpriors and priors are also performed to check the system identifiability for both applications.

2.7 Parameter sensitivity analysis

After the unknown parameters have been estimated for the given the model, the next step is to assess the sensitivity of these parameters. Parameter sensitivity analysis is performed to investigate the effect of change in a parameter value on the overall system. It provides critical information on the relationship between parameters and system outputs. With parameter sensitivity analysis, it can be assessed whether the system is robust enough to operate reliably when its parameters vary within their expected ranges. A small variation in a highly sensitive parameter may introduce fragility into the system [103]. On the other hand, if the clinical outcome of interest is sensitive to one or several parameters in the system, the parameter may serve as an indicator of the clinical outcome. There are two categories of sensitivity analysis techniques: local sensitivity analysis and global sensitivity analysis [104].

One of the simplest local sensitivity analysis methods is one-at-a-time analysis. The parameter sensitivity can be measured by repeatedly varying the parameter of interest while holding the other parameters fixed. The sensitivity of a parameter θ_i can be reflected by the change in the system output towards the percentage change in that parameter [103]. It is easy to measure the change in the system outcome when the output is one single value; however, the output in our applications is a time series. Therefore, a function of the output W is defined to quantify the change in the output

time series, and δW is the change in W caused by the change in the parameter $\delta\theta_i$. Typically W are: the sum of the squared differences between the output produced by a reference parameter and the perturbed system output [105], the area under the curve of the output [106], the amplitude and period of oscillation [107], etc. In this thesis, since root mean square is one of the most widely used measures of model goodness-of-fit, W is defined as the root mean square difference between the measurements and the output produced by the deterministic parameters with a small perturbation in θ_i :

$$W = \text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^T (\tilde{y}_t - y_t)^2} \quad (2.82)$$

where y_t is the measurement value at time t , \tilde{y}_t is generated using the mean values of the deterministic parameters, denoted as μ_{θ_i} , with a small perturbation. A reference $W^{(0)}$ can be obtained when the mean values of the deterministic parameters are used without perturbation. To test the sensitivity of the parameter θ_i , 1000 samples of θ_i are drawn from a uniform distribution from $0.99\mu_{\theta_i}$ to $1.01\mu_{\theta_i}$, each sample $\tilde{\theta}_i^{(j)}$ ($j = 1, 2, \dots, 1000$), along with other parameters $\theta_{\setminus i}$, are used to generate a deterministic time series that deviates from the time series generated by θ . The change of percentage in W compared with the reference value $W^{(0)}$ for each time series is then divided by the change of percentage in θ_i accordingly. The sensitivity of the parameter θ_i , denoted as the sensitivity index SI_{θ_i} , is therefore defined as follows:

$$SI_{\theta_i} = \frac{1}{1000} \sum_{j=1}^{1000} \frac{\delta W^{(j)} / W^{(0)}}{\delta \theta_i^{(j)} / \mu_{\theta_i}} \quad (2.83)$$

The sensitivity index does not have a unit, since it represents a ratio of change in an output to a change in a parameter. This is a local sensitivity analysis, because it only considers the sensitivity of the parameters around a certain chosen point of the parameter space. For the VB method in this thesis, all the inferred parameters follow a normal distribution, therefore, a small value in the RMSE between the measurement points and the output generated by the mean values of the parameters, cannot guarantee small RMSE values generated with parameter values within a small neighbourhood of the mean values. To guarantee the robustness of the model with the inferred parameter values with a level of uncertainties, it is important to check the absolute value of the RMSE and the sensitivity index.

The other group of sensitivity analysis techniques not only considers the individual parameter sensitivity, but also assesses the combined variability resulting from considering all parameters simultaneously. Random samples of parameters in the whole parameter space, instead of one parameter, are generated based on the inferred posterior distributions. The change in the output towards the change in the parameter can be used to assess the parameter sensitivity in a ‘global’ sense. In this way, the obtained parameter sensitivity is dependent on the interactions and influences of all of the parameters. The drawback of global sensitivity analysis is the high computational cost.

In this thesis, one-at-a-time analysis is performed to assess the sensitivity of model parameters in Chapter 3 and Chapter 4.

2.8 Illustrative example of the use of the VB toolbox

In this section, an example is given to show how the VB method works given a single measurement time series, and how different priors and hyperpriors influence the free energy. A VB toolbox [77], initially developed for neuroimaging data, is applied to perform the inference. The toolbox was designed to provide a flexible platform to deal with nonlinear dynamic models in continuous time. It is capable of performing efficient and robust parameter estimation, and providing quantitative diagnostics of model fitting. This section presents the parameter inference and model selection procedure for a toy model based on second order SDEs. A time series was generated using the toy model and the strategy of obtaining the parameters and selecting the best model has been investigated. This strategy is then applied to real data in later chapters.

The toy model was chosen in the form of (2.2). The VB method was then applied to infer the model parameters and choose the best model from several model candidates. Tests were performed to see whether the program can select the right structure and infer the parameters successfully.

2.8.1 The toy model

The toy model is linear and is based on (2.2), in which $n = 2$, $f_1 = \theta_1 \dot{x}_t$, $f_0 = \theta_2 x_t + \theta_3$, $u_t = 0$:

$$\begin{cases} \ddot{x}_t + \theta_1 \dot{x}_t + \theta_2 x_t + \theta_3 = \eta_t \\ y_t = x_t + \varepsilon_t \end{cases} \quad (2.84)$$

with the initial condition: $\mathbf{x}_0 = \begin{pmatrix} x_0 \\ \dot{x}_0 \end{pmatrix}$

In (2.84), η_t corresponds to the system noise and ε_t represents the measurement noise, both of which are modelled as AWGN noise: $\eta_t \sim N(0, \alpha_1^{-1})$ and $\varepsilon_t \sim N(0, \alpha_2^{-1})$; α_1 is the precision of the dynamic noise and α_2 is the precision of the measurement noise. The toolbox is capable of accounting for either measurement noise only or for both system and measurement noise, with two sets of inference results accordingly. To differentiate these two sets of inference results, in this thesis, the model with measurement noise only is referred to as the *deterministic model*, and the model with both forms of noise is referred to as the *stochastic model*.

The system equation of (2.84) can be written in state-space form as:

$$\begin{pmatrix} x_t \\ \dot{x}_t \end{pmatrix}' = \begin{pmatrix} 0 & 1 \\ -\theta_2 - \theta_3/x_t & -\theta_1 \end{pmatrix} \begin{pmatrix} x_t \\ \dot{x}_t \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \eta_t \quad (2.85)$$

Denote $\mathbf{x}_t = \begin{pmatrix} x_t \\ \dot{x}_t \end{pmatrix}$ as the state vector, a set of variables representing the configuration of the system. The system reaches its steady state when $\mathbf{x}_{steady} = \begin{pmatrix} -\theta_3/\theta_2 \\ 0 \end{pmatrix}$. A time series \mathbf{y} consisting of 100 time points with the time interval of $\Delta t = 1$ (Fig. 2.4) was simulated using the parameter values listed in the second column of Table 2.2. In the simulated time series, both noise intensities, α_1^{-1} and α_2^{-1} , were set to be 0.01, which was relatively small compared with the peak height around 350.

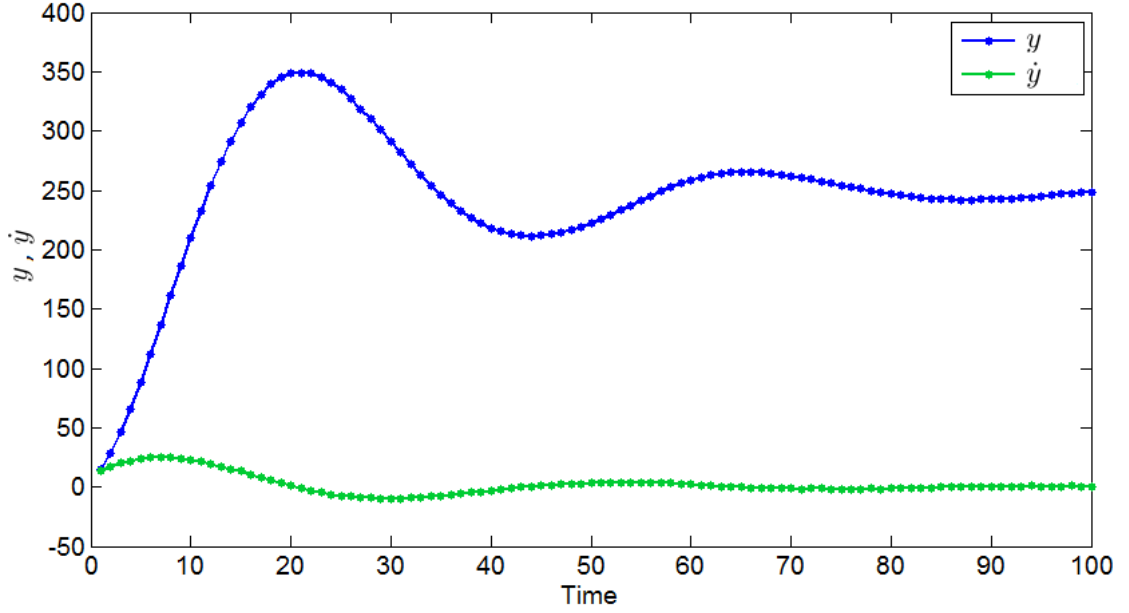


FIGURE 2.4: Simulated time series from the toy model.

TABLE 2.2: Summary of the parameter settings for the time series simulated by the toy model

Parameter	Values for simulation	Prior
θ	$\begin{pmatrix} 0.1 \\ 0.02 \\ -5 \end{pmatrix}$	$\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10^4 & 0 & 0 \\ 0 & 10^4 & 0 \\ 0 & 0 & 10^4 \end{pmatrix}\right)$
α_1	100	$\mathcal{Ga}(0.001, 0.001)$
α_2	100	$\mathcal{Ga}(0.001, 0.001)$
x_0	$\begin{pmatrix} 5 \\ 10 \end{pmatrix}$	$\mathcal{N}\left(\begin{pmatrix} x_0 \\ \dot{x}_0 \end{pmatrix}, \begin{pmatrix} 10^4 & 0 \\ 0 & 10^4 \end{pmatrix}\right)$

With the simulated time series, parameter inference was performed using the VB toolbox and the inferred parameters were compared with the parameter values used for simulation. Note that this simulated time series has a relatively large measurement noise precision, which is 100. When the simulated measurement noise precision decreases, the uncertainty of the inferred parameters and hyperparameters may increase.

2.8.2 The choice of priors

This section is devoted to illustrating how the priors/hyperpriors for the parameters/hyperparameters were selected for the toy model. The VB method has been applied

with different prior settings to infer the parameters and the hyperparameters of the simulated time series. The following prior information was required: the priors for the deterministic parameters, the hyperpriors for the stochastic parameters, and the priors for the initial conditions. To compare the inference results between the deterministic model and the stochastic model, all the priors for the deterministic model, except for the system noise hyperpriors, were set the same as for the stochastic model. All these priors can influence the inferred posterior distributions of the parameters and the value of the free energy, as shown in Section 2.3.4. As a larger free energy value indicates a better model, the priors/hyperpriors of the parameters/hyperparameters set by the modeller are important for model comparison. The influence of prior over posterior can be viewed in two ways. On the negative side, the dependence of free energy on priors requires exploration over the parameter space to achieve the optimal free energy for each model candidate. On the positive side, a better knowledge of the parameter range can constrain the parameter space that needs to be explored. Therefore, thorough exploration is only needed for some typical time series, and the selected prior can be applied to similar time series.

To illustrate the importance of prior settings, they were initially set as the values shown in Table 2.2. To quantify the influence of the priors over the value of free energy, various prior settings have been explored to provide an insight into how to select the priors wisely for different parameters.

2.8.2.1 Initial default setting of priors and hyperpriors

As discussed in Section 2.5, weakly informative priors were initially set for all of the parameters to make sure that the prior would not overpower the data. For the deterministic parameters modelled as Gaussian distributions, a typical weak prior was set with zero mean and relatively large variances as shown in the third column of Table 2.2. The variances, set to be 10^4 , allowed the algorithm to search over a relatively wide region for the optimal posterior distributions for the parameters. Both noise precisions, α_1 and α_2 , were modelled by gamma distributions: $\alpha_i \sim \mathcal{Ga}(a_i, b_i)$ ($i = 1, 2$). Both hyperparameters a_i and b_i ($i=1,2$) were set to 0.001, which is a typical weakly informative prior (see detail in Section 2.5).

The free energy value for the stochastic model, denoted as \mathcal{F}_s , is -248 , and the free energy value for the deterministic model, denoted as \mathcal{F}_d , is -1059 . The free energy is the lower bound of the log-evidence of the model, $\log P(\mathbf{y}|M)$. The higher free energy for the stochastic model indicates that the stochastic model is more probable. The inferred posterior distribution of the parameters follows a normal distribution as $\boldsymbol{\theta} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \Sigma_{\hat{\boldsymbol{\theta}}})$, where $\hat{\boldsymbol{\theta}}$ is the vector of the posterior mean values of the parameters and $\Sigma_{\hat{\boldsymbol{\theta}}}$ is the posterior covariance matrix of the parameters:

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} 0.104 \\ 0.022 \\ -5.53 \end{pmatrix} \quad \Sigma_{\hat{\boldsymbol{\theta}}} = \begin{pmatrix} 7.98 \times 10^{-5} & 5.63 \times 10^{-6} & -0.0016 \\ 5.63 \times 10^{-6} & 1.35 \times 10^{-6} & -3.39 \times 10^{-4} \\ -0.0016 & -3.39 \times 10^{-4} & 0.09 \end{pmatrix} \quad (2.86)$$

The posterior hyperparameters for the system noise, \hat{a}_1 and \hat{b}_1 , and the posterior hyperparameters for the measurement noise, \hat{a}_2 and \hat{b}_2 , are as follows:

$$\hat{a}_1 = 100, \quad \hat{b}_1 = 41, \quad \hat{a}_2 = 100, \quad \hat{b}_2 = 38. \quad (2.87)$$

The inferred posterior distribution of the initial condition follows a normal distribution as $\mathbf{x}_0 \sim \mathcal{N}(\hat{\mathbf{x}}_0, \Sigma_{\hat{\mathbf{x}}_0})$, where $\hat{\mathbf{x}}_0$ is the vector of the means of the initial condition and $\Sigma_{\hat{\mathbf{x}}_0}$ is the covariance matrix of the initial condition:

$$\hat{\mathbf{x}}_0 = \begin{pmatrix} 7.38 \\ 6.24 \end{pmatrix} \quad \Sigma_{\hat{\mathbf{x}}_0} = \begin{pmatrix} 0.88 & -0.48 \\ -0.48 & 0.49 \end{pmatrix} \quad (2.88)$$

Compared with the parameter values used for the simulation as seen in Table 2.2, the estimation errors (difference between the mean of the posterior and the true value) for the parameters θ_i ($i=1, 2, 3$) are 0.004, 0.002 and -0.53 receptively. The correlation between the parameters is shown in Fig. 2.5, from which it can be seen that the correlation between θ_2 and θ_3 is high (Fig. 2.5). The relative higher estimation errors in θ_2 and θ_3 (around 10%) compared with the estimation error in θ_1 (4%) can be explained by the high correlations in the covariance matrix between θ_2 and θ_3 (see the caption of Fig. 2.5), which causes difficulties in identifying the parameters individually as explained in Section 2.3.4. The mean precisions of the posterior system noise and measurement noise are $\frac{\hat{a}_1}{\hat{b}_1} = 2.44$ and $\frac{\hat{a}_2}{\hat{b}_2} = 2.63$, both of which greatly underestimate the real precision of 100. This underestimation is caused by poor selection of the priors and the hyperpriors. Based

on the hyperpriors, the prior mean for the both noise intensities is 1. The posterior mean for both noise intensities has increased to larger than 2 in light of the data. However, the influence from the prior over the posterior is not negligible, and this causes the underestimation of both noise intensity.

By exploring different settings for the prior/hyperpriors of the parameters/hyperparameters, the estimation errors of the parameters/hyperparameters can be further decreased and the free energy value can be further increased, implying that the gap between the free energy and the log-evidence for the model can be further decreased.

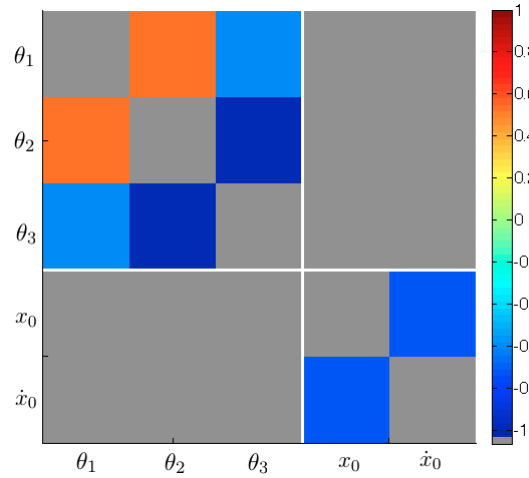


FIGURE 2.5: The correlations between the parameters θ_1 , θ_2 , θ_3 and the initial conditions x_0 , \dot{x}_0 . The correlation between θ_2 and θ_3 is 0.97, between θ_1 and θ_2 is 0.54, and between θ_1 and θ_3 is 0.60.

2.8.2.2 Select the mean vector of the prior distribution for the parameters

To see if the inferred posterior distributions of the parameters remain unchanged with different mean vectors of the prior distributions for the parameters, the prior means of each parameter ranging from -30 to 30 have been used to obtain the posteriors.

First, with the priors of θ_1 and θ_2 fixed, the prior mean of θ_3 was changed ranging from -30 to 30 with an increment of 1. The free energy values and the posterior mean $\hat{\theta}_3$ of θ_3 are shown in Fig. 2.6. When the prior values are around zero, the change of the free energy is dramatic. The deviation from the true value of θ when the prior mean is zero might be caused by multi-modality in the model evidence. Therefore, the prior mean should be chosen outside of the range from -1 to 1 .

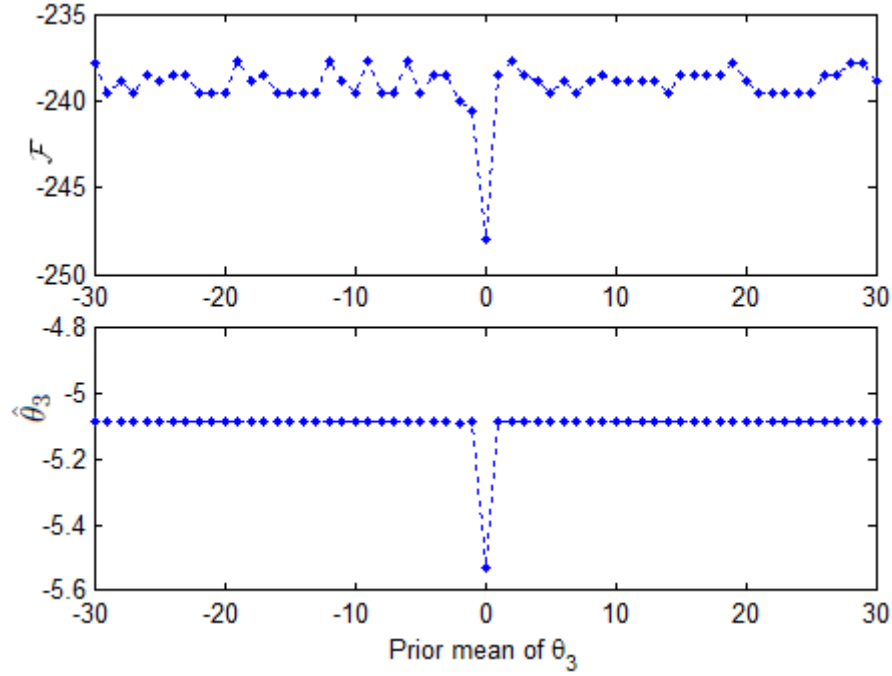


FIGURE 2.6: Free energy and posterior means of θ_3 using different prior means of θ_3 .

Second, with the priors of θ_1 and θ_3 fixed, the prior mean of θ_2 was changed from -30 to 30 , and the corresponding inferred posterior mean $\hat{\theta}_2$ is shown in Fig. 2.7. It shows that changing the prior mean of θ_2 can cause dramatic changes in the free energy and the posterior mean. From the bottom figure in Fig. 2.7, the posterior mean $\hat{\theta}_2$ increases when the prior mean increases except when the prior was set between -3 and 3 . But compared with the top figure in Fig. 2.7, the free energy dramatically changes in that region, and the prior setting that generates the highest value of free energy is when $\theta_2 = 0$. The possible explanation is that the parameter θ_2 might only be locally identifiable, and with a prior that is far away from its true value, the posterior mean will be determined by the prior mean. The sensitivity of the free energy to the prior mean near the true value of the mean makes it necessary to explore different prior means for θ_2 to select the appropriate parameter prior.

Finally, the priors of θ_2 and θ_3 were fixed, and the prior mean $\hat{\theta}_1$ was changed from -30 to 30 as shown in Fig. 2.8. When the prior mean is larger than 27 , which is far from the true value of 0.1 , the posterior mean $\hat{\theta}_1$ settles at around 0.2 . However, from the much smaller value of the free energy, the inferred value is not optimal. From the magnified box shown in Fig. 2.8, different prior means with a smaller step of 0.1 between -1 and 1 were used. The posterior mean $\hat{\theta}_1$ is not as sensitive to the prior as θ_2 and θ_3 , but a

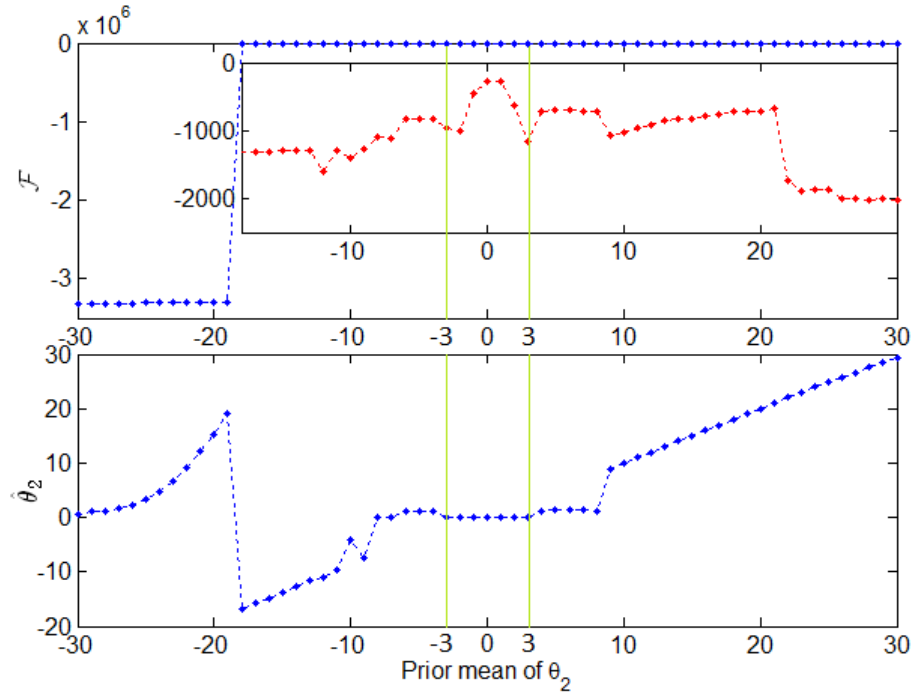


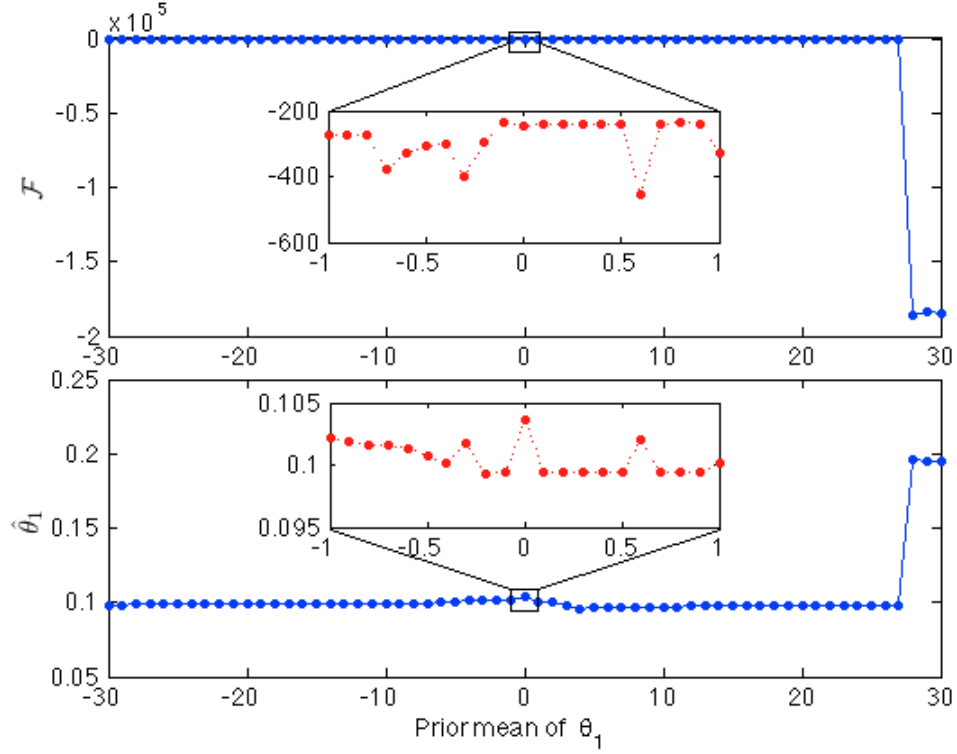
FIGURE 2.7: Free energy and posterior means of θ_2 using different prior mean values of θ_2

small change in the prior mean can still cause a big difference in the free energy values.

It is clear that the change of prior means of the three parameters have different levels of influence over the posterior means and the free energy values. In fact, the degree of influence of the prior mean of the parameters also depends on the number of the measurement data points. As explained in Section 2.5, the posterior distribution of the parameters is dependent on both the priors and the data, which means more weight would be put on the priors when there are less measurement data. Thus, to make sure the inferred posterior is robust with respect to different prior settings, different prior means of the parameters should be considered by covering a reasonable area in the parameter space in order to find the ‘true’ value of the parameters as indicated by the free energy.

2.8.2.3 Select the variances of prior distributions for the parameters

The posteriors of the parameters are influenced not only by the mean value of the prior distributions, but also by the variances of the prior distributions. In this toy model, the covariance matrix of the prior distributions for θ_1 , θ_2 and θ_3 are set to be $10^4 \mathbb{I}_3$ (\mathbb{I}_3 is the 3×3 identity matrix). The large variance allows a wide range of parameter values to be

FIGURE 2.8: Free energy and posterior means of θ_1 using different prior means of θ_1

explored in parameter space, and also expresses the uncertainty on the parameter values before the data are observed. If certain information of the parameter value is known, then the variance of the parameter can be set to a narrower range to incorporate the known information, which may lead to a smaller variance in the posterior distribution, and improve the estimation of the parameters. To explore the influence of the variances of the parameter prior distributions for the parameters, the variances of the priors $10^j \mathbb{I}_3$ were changed from $j = -2$ to $j = 8$. The free energy for different settings is shown in Fig. 2.9. To quantify the estimation errors for all three parameters, the *Root Mean Square Error* (RMSE) between the real values and the estimated mean values ($\hat{\theta}_i$, $i = 1, 2, 3$) is calculated as follows:

$$\text{RMSE}_\Sigma = \sqrt{\frac{(\theta_1 - \hat{\theta}_1)^2 + (\theta_2 - \hat{\theta}_2)^2 + (\theta_3 - \hat{\theta}_3)^2}{3}} \quad (2.89)$$

As shown on the bottom graph in Fig. 2.9, when the covariance matrix of the prior distribution is set to $10 \mathbb{I}_3$, i.e. $j = 1$, the estimation errors of the parameters are the smallest with RMSE_Σ of 0.019 and free energy of -230 . From Fig. 2.9, when the variance of the prior distribution is set too small ($j < 0$), the area in the parameter space that can be explored is highly limited, and therefore, the free energy \mathcal{F} settles at a suboptimal

value. When the variance of the prior distribution are increased from $j = -2$ to $j = 0$, the estimation error of the parameters between the posterior mean and the true value decreases, and remains low when the variances of the priors are further increased from $j = 0$ to $j = 8$.

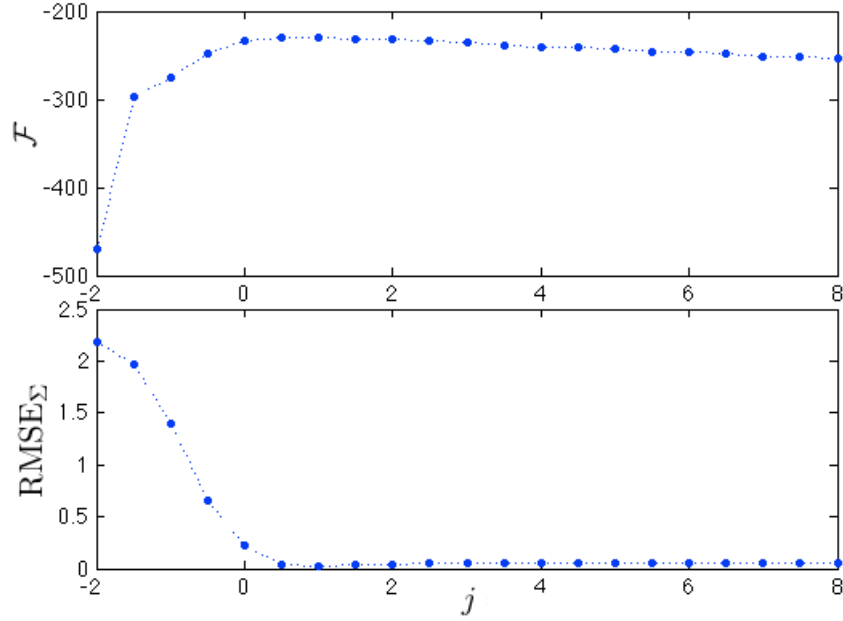


FIGURE 2.9: Free energy \mathcal{F} and RMSE_{Σ} with the variance of the prior distribution for all three parameters $10^j \mathbb{I}_3$ changing from $j = -2$ to $j = 8$.

For this example, it is clear that when the variances of the parameters are larger than 10, the accuracy of the parameter inference stays the same and the slight difference in free energy is only caused by the different priors. Therefore, the priors for the variances of the parameter shows robustness when large prior settings for the variances are chosen (≥ 10), and the posterior distributions of the parameters do not rely on the prior variance for the parameters. To conclude, it is better to set a larger variance as a weak prior in real applications.

2.8.2.4 Select the hyperpriors for the hyperparameters

The precisions of both forms of noise are modelled by Gamma distributions: $f(\alpha_i) \sim \mathcal{G}(a_i, b_i)$ where a_i is the shape parameter and b_i is the rate parameter. Correspondingly, the probability density function is as follows:

$$f(\alpha_i) = \frac{b_i^{a_i}}{\Gamma(a_i)} \alpha_i^{a_i-1} e^{-b_i \alpha_i} \quad (2.90)$$

An illustrative graph is shown in Fig. 2.10 with different values of a_i and b_i where a_i/b_i is the mean of the distribution, and a_i/b_i^2 is the variance.

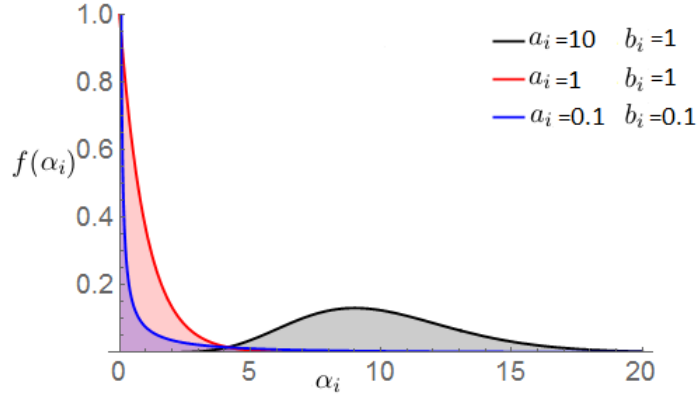


FIGURE 2.10: Probability density distribution of the precision of the noise

Let us consider how the hyperpriors a_i and b_i in both types of noise (system noise and measurement noise) influence the posterior distribution of the parameters. As explained in Section 2.5, setting all the hyperparameters (a_1, a_2, b_1, b_2) to 0.001 can provide weakly informative hyperpriors for the precision of both forms of noise. Different combinations of the hyperprior setting have been used to study the influence of the hyperpriors over the estimation error for the noise intensity and the value of the free energy. The estimation error for both forms of noise intensity can be calculated as follows:

$$\text{RMSE}_\alpha = \sqrt{\frac{(\frac{\hat{a}_1}{\hat{b}_1} - 100)^2 + (\frac{\hat{a}_2}{\hat{b}_2} - 100)^2}{2}} \quad (2.91)$$

First, fix the hyperpriors for the system noise, a_1 and b_1 , to 0.001, and consider the influence of the hyperpriors of the measurement noise over the free energy value and the posterior distributions of both forms of noise intensity. Varying the values of a_2 and b_2 results in different values of free energy as shown in Fig. 2.11 (a). It is clear from the figure that the free energy values along the diagonal lines (parallel to the indicated line) are close. The free energy values are maximised along the indicated diagonal line $\log(a_2) = \log(b_2) + 2$, where the prior mean is $a_2/b_2 = 100$. This means that the prior mean of the measurement noise intensity has a non-negligible influence over the free energy values. The RMSE between the estimated mean of the noise precision and the true noise precision was calculated according to (2.91). The logarithm of the RMSE_α ($\log \text{RMSE}_\alpha$) for different combinations of the shape (a_2) and rate (b_2) hyperparameters

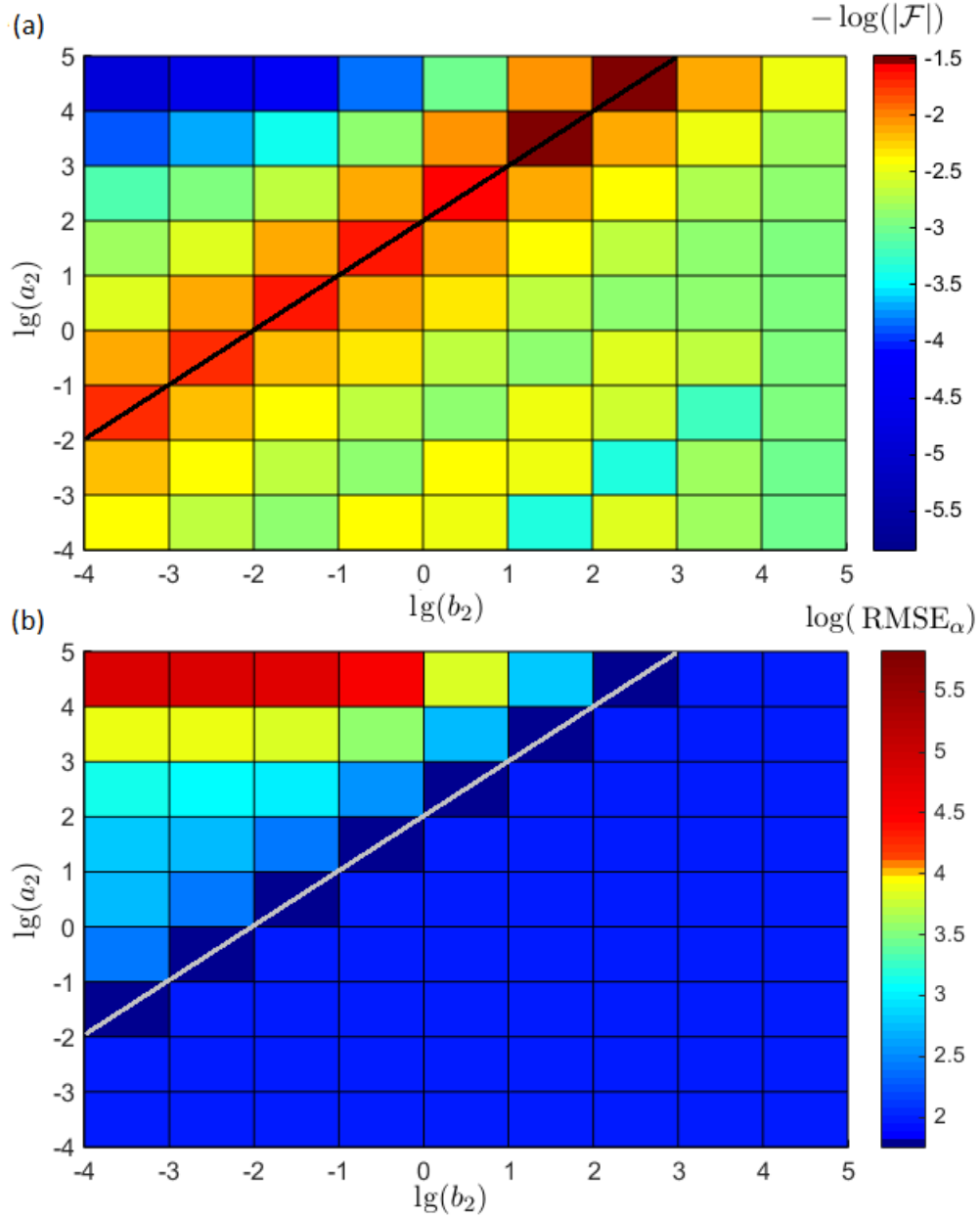


FIGURE 2.11: (a) The value of free energy and (b) RMSE_α for different combinations of the shape a_2 and rate b_2 hyperpriors for the measurement noise. Note that the figure is shown in log-scale (with a base of 10). All the free energy values \mathcal{F} in this graph are negative, so the logarithm of the free energy is calculated by $-\lg(|\mathcal{F}|)$.

of the measurement noise is shown in Fig. 2.11 (b). These results agree with the free energy results shown in Fig. 2.11 (a): when the prior mean of the measurement noise precision a_2/b_2 is set to 100, \mathcal{F} is maximised and RMSE_α is minimised. Along the indicated diagonal line in both Fig. 2.11 (a) and Fig. 2.11 (b), the variance of the prior distribution for the measurement noise precision, a_2/b_2^2 , decreases when $\log(b_2)$ increases. To make sure that the variance of the prior distribution of the noise intensity is large

enough, the variance of the prior distribution of the noise intensity should be at least 10^2 , which means that b_2 should be no larger than 1. When b_2 is set to 1, and a_2 is 10^2 , the posterior estimations for the hyperparameters are: $\hat{a}_1 = 100.001$, $\hat{b}_1 = 5.31$, $\hat{a}_2 = 200$, $\hat{b}_2 = 2.22$, and the free energy is -21 , which is the largest value. It is worth noticing that the posterior mean of the system noise precision $\hat{a}_1/\hat{b}_1 = 19$, and the posterior mean of the measurement noise precision is $\hat{a}_2/\hat{b}_2 = 90$. Both estimated precisions are smaller than the real precisions of 100, with the measurement noise precision closer to the real value. The underestimated noise precision indicates that the VB method tends to be conservative in terms of noise precision inference.

Second, fixing the hyperpriors for the measurement noise, a_2 and b_2 , to 0.001, and consider the influence of the hyperpriors of the system noise over the free energy value and the posterior distribution of both forms of noise intensity. Varying the value of a_1 and b_1 from 10^{-4} to 10^5 (order 10 incremental) results in different values of the free energy \mathcal{F} as shown in Fig. 2.12 (a) and the RMSE_α as shown in Fig. 2.12 (b). Note that varying the value of a_1 and b_1 from 10^{-4} to 10^5 has covered a huge range of possible noise intensities from 10^{-9} to 10^9 , which is considered sufficiently wide. Similar to the mean of the prior distribution of the measurement noise intensity, the mean of the prior distribution of the system noise, a_1/b_1 , also influences the value of \mathcal{F} and the RMSE_α . It is worth noticing the high free energy values on the top left corner of Fig. 2.12 (a). Take $a_1 = 10$ and $b_1 = 10^{-4}$ for example, the free energy value, $\mathcal{F} = 473$, and is large. This indicates that the probability of the model being true with the inferred parameters is high. However, from Fig. 2.12 (b), the RMSE_α with hyperpriors $a_1 = 10^{-4}$ and $b_1 = 10^1$ is as high as 7.8×10^5 , indicating an inaccurate inference. The contradictory result of a high free energy and a high RMSE_α is caused by the bad choice of the hyperpriors of the system noise with a mean value of $a_1/b_1 = 10^{-5}$ and a variance of $a_1/b_1^2 = 10^{-6}$. With such a strong informative hyperprior, the posterior hyperparameters $\hat{a}_1 = 110$, $\hat{b}_1 = 10^{-4}$ are completely dominated by the prior distribution. The artificial high confidence level of the small system noise precision boosts up and dominates the value of the free energy. Therefore, the mean of the prior a_1/b_1 is held at 100 indicated by the straight line shown in Fig. 2.12 (a). Finally, $a_1 = 100$ and $b_1 = 1$ are selected based on the free energy value of 2. It is the largest among all of the combinations of the hyperparameters except for the top left corner where the free energy values do not reflect the goodness of fit. The posterior hyperparameter values are $\hat{a}_1 = 200$, $\hat{b}_1 = 1.54$, $\hat{a}_2 = 100.001$, $\hat{b}_2 = 10.30$.

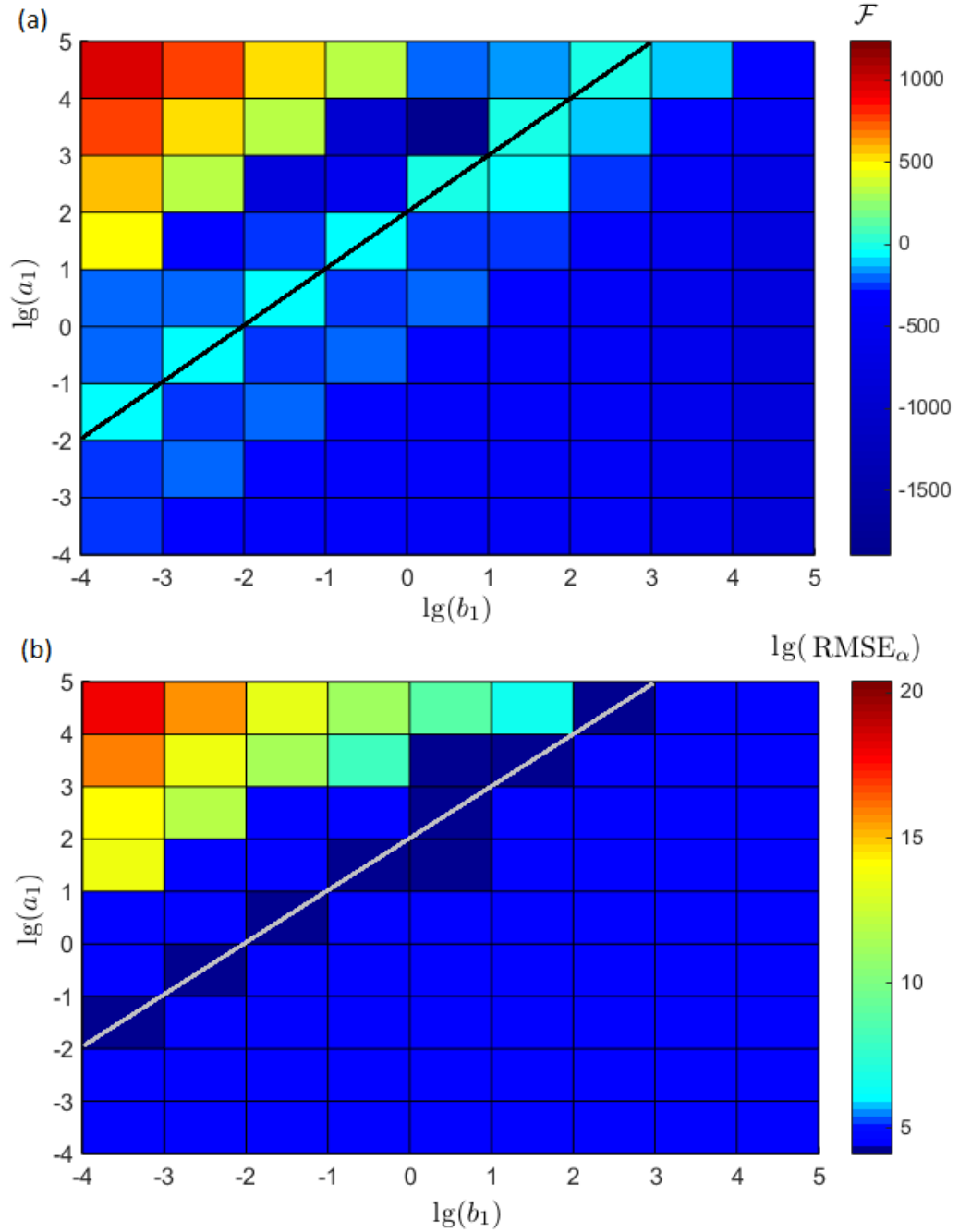


FIGURE 2.12: (a) The value of free energy and (b) RMSE_α for different combinations of the shape a_2 and rate b_2 hyperpriors for the system noise. Note that the figure is shown in log-scale (with a base of 10).

The posterior mean of the system noise is $\hat{a}_1/\hat{b}_1 = 130$, which is overestimated and the posterior mean of the measurement noise is $\hat{a}_1/\hat{b}_1 = 9.71$, which is underestimated.

The optimal free energies of the second step (fixing the hyperpriors for the measurement noise) are higher than for the first step (fixing the hyperpriors for the system noise). Therefore, the hyperpriors are finally chosen as: $a_1 = 100$, $b_1 = 1$, $a_2 = b_2 = 0.001$. The

mean of the posterior distribution for the parameters θ and the covariance matrix $\Sigma_{\hat{\theta}}$ are as follows:

$$\hat{\theta} = \begin{pmatrix} 0.099 \\ 0.020 \\ -5.03 \end{pmatrix} \quad \Sigma_{\hat{\theta}} = \begin{pmatrix} 1.53 \times 10^{-6} & 1.13 \times 10^{-7} & -3.11 \times 10^{-5} \\ 1.13 \times 10^{-7} & 2.60 \times 10^{-8} & -6.58 \times 10^{-6} \\ -3.11 \times 10^{-5} & -6.58 \times 10^{-6} & 0.0017 \end{pmatrix} \quad (2.92)$$

The posterior hyperparameters for the system and measurement noise are as follows:

$$\hat{a}_1 = 200 \quad \hat{b}_1 = 1.54 \quad \hat{a}_2 = 100.001 \quad \hat{b}_2 = 10.30 \quad (2.93)$$

Compared with the posterior distributions of the parameters obtained using well-selected priors with the posterior result shown with (2.86) and (2.87) obtained using the default priors, every inferred parameter is closer to its real value as shown in Table 2.2. Therefore, it can be concluded that selecting the priors and hyperpriors would improve the performance of the algorithm.

It is worth noticing that the improvement of the inference is reflected not only by the improved mean values of the parameters, but also the smaller variances of the parameter posterior distributions and a smaller error of the estimation for both forms of noise intensity. The free energy, increased from -284 to 2 , proves to be a good indicator of selecting priors and hyperpriors, except when the prior variances are set too narrow. When the VB method is applied in practice and the real values of the parameters are unknown, not only a high free energy value, but also a small estimated variance of the parameter distributions, and a high estimated noise precision indicate a good choice of the priors and hyperpriors for the parameters and hyperparameters.

2.9 Chapter summary

The methodology that is later applied in Chapters 3 and 4 has been presented in this chapter. The model framework based on differential equations has been presented in Section 2.1, and this framework is going to be used for model development for two medical applications in Sections 3.3.1 and 4.3. The VB method has been described in detail in this chapter, and its advantages and disadvantages compared to other sampling methods such as MCMC and SMC have also been discussed. Free energy calculated

using the VB method serves as a model selection criterion, and it has been compared to other criteria such as the AIC and BIC. It has been stressed that the VB method, as a full Bayes approach, provides a probabilistic view towards every unknown variable including the parameters, the hyperparameters, the initial conditions and the states, and the calculation of free energy takes the uncertainties of these unknown variables into consideration. The VB method will serve as the main tool to select the best model structure and learn the parameter values in applications presented in Sections 3.3.2 and 4.3.2. The choices of priors and hyperpriors for parameters and hyperparameters have been discussed and their influence over the posteriors has been highlighted in the provided example. The specific choice of priors will be further investigated in Sections 3.4.1 and 4.3.2.

The format and the quality of the data collected in the two clinical applications are different. Both sets of measurement data consist of one single time series. The measurement data collected for the first application are sampled with equal spacing without any missing data. The measurement data collected for the second application are irregularly sampled, and the data are relatively sparse. However, the VB method is successfully employed to select the best model that describes the data for both applications, demonstrating the flexibility and capability of the method.

Chapter 3

Post-prandial glucose dynamics

With an ever increasing population with Diabetes Mellitus (DM), scientists have been endeavouring to establish and develop various mathematical models to describe or predict in human glucose dynamics. This chapter presents a data-driven model that has been developed using the VB method for model selection and parameter inference. Inspired by a linear deterministic model built by Wu [108] based on data from one DM patient, a stochastic model with a second order nonlinear differential equation has been developed to describe the response of blood glucose concentration to food intake using continuous glucose monitoring data for people with and without DM. The number and values of the system parameters were defined by iterative optimisation of the free energy as introduced in Chapter 2. A comprehensive analysis demonstrated that deterministic system parameters belong to different ranges for DM and non-DM profiles. Implications for clinical practice are discussed. This is the first study introducing a continuous data-driven nonlinear stochastic model capable of describing blood glucose dynamics for people with and without DM. This work has been published in [109, 110]. The structure of the chapter is as follows. Section 3.1 gives a background introduction to DM and the glucose regulatory system, and discusses the models available in the literature. Section 3.2 provides details on the data available for the analysis. Section 3.3 presents the process of the model formulation and development, and explains the VB method that has been adapted to this application. Detailed analysis of system parameters and comparison of the parameters between the groups of people with and without DM are discussed in Section 3.4 together with clinically related interpretations. Section 3.5 summarises the results.

3.1 Introduction

Diabetes Mellitus (DM), commonly referred to as diabetes, poses an alarming threat in modern public health with rising trends and severity in recent years. According to an International Diabetes Federation report [111], an estimated 387 million people have DM worldwide, and the number is expected to rise to 592 million by 2035. DM is a group of diseases characterised by high blood glucose (known as hyperglycemia) levels that result from defects in the body's ability to produce and/or use insulin. The chronic hyperglycemia of DM can lead to various microvascular and macrovascular complications, including blindness, limb loss, ischemic heart disease and end-stage renal disease [112]. The majority of the DM cases fall into two categories, both of which result from complex interactions between the genes and the environment, but the pathophysiology and the treatments are different.

Type 1 diabetes (T1D), which accounts for 5 – 10 % of the diabetes population, usually starts in childhood and adolescence. It is caused by the inability of the β cells in the pancreas to produce insulin. Therefore, T1D patients need to be treated with insulin injection or insulin pumps to control hyperglycemia. Poor judgement of the amount of the insulin injected can result in serious hypoglycemia (low blood glucose) which may lead to brain damage, coma and even death [113].

Type 2 diabetes (T2D), which accounts for 90 – 95 % of the diabetes population, is caused by a combination of tissue resistance to insulin and an inadequate compensatory insulin secretion. T2D occurs more frequently with increasing age, and is highly related with a sedentary modern lifestyle with excess caloric intake. Various oral anti-diabetic drugs [114] that reduce insulin resistance or increase insulin sensitivity, as well as insulin, are used for treatments. Lifestyle changes, such as limitation of energy intake, regular exercise, and weight reduction, have also proved beneficial to control [115, 116] and even reverse the progression of the disease [117]. T2D has long been considered as a chronic progressive condition and has several different stages [118]. It may stay undetected for a long period of time without clinical symptoms when the underlying disease is developing and causing damage in β cells.

When the glucose level is higher than normal, but not yet high enough to be classified as T2D, it is usually referred to as pre-DM. Without intervention, people with pre-DM have

a higher risk of progressing to T2D; thus there is an urgent need for improved diagnostic methods that are capable of detecting symptoms at early stages of the disease. For people with DM or in the process of progressing into DM, it is crucial to achieve the goal of maintaining *glycaemic stability* during daily life, which requires a deep understanding of which and how different factors (food intake, exercise, mental stress, etc.) influence the glucose variations [119].

This chapter is focused on one important factor – food intake – that influences glucose variations in daily lives. Each food intake is considered as one *event* that serves as an external excitation force to the glucose regulatory system. The glucose excursions afterwards reflect the *transient response* of the system to such excitation, and eventually the glucose concentration settles back to the baseline, which represents the *steady state* of the system.

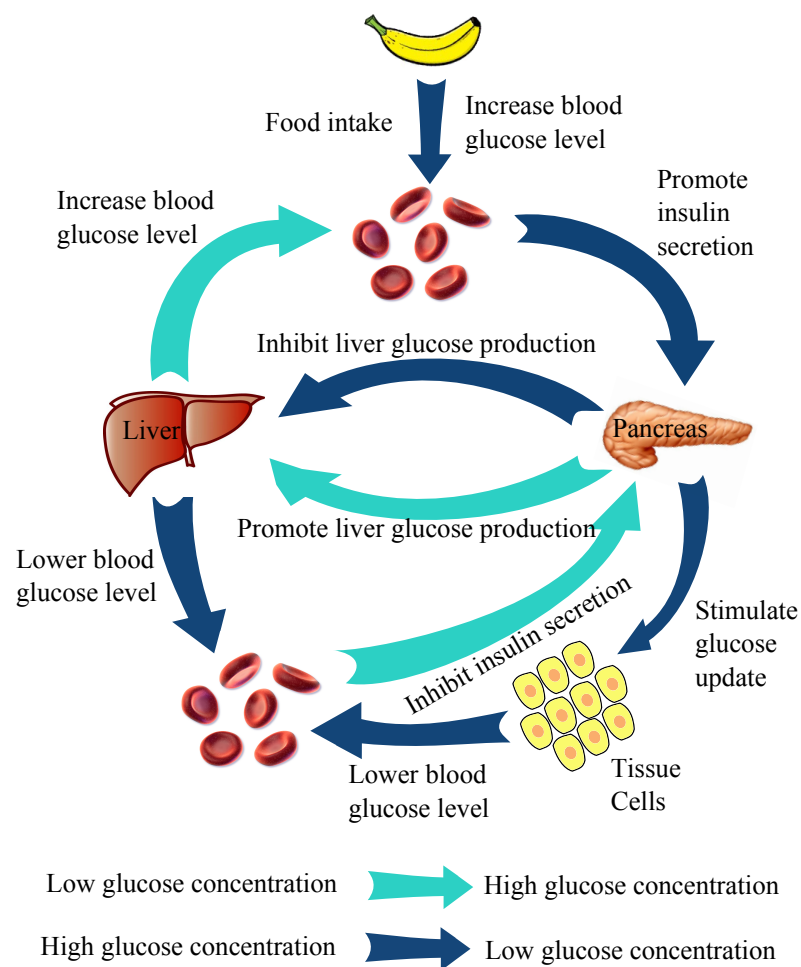


FIGURE 3.1: Simple illustration of the glucose - insulin feedback system.

A wide variety of hormones take part in glucose regulation including insulin, glucagon,

epinephrine, cortisol and growth hormone. Among them, the main regulator is insulin. As indicated by the dark blue arrows in Fig. 3.1, when the glucose perturbation occurs (after a meal), β cells in the pancreas secrete more insulin in response to the excessive glucose and the increased insulin promotes the glucose utilisation in various tissues and inhibits the hepatic glucose production to bring the blood glucose effectively down to the pre-perturbation level. Note that the blood in the human body normally only contains less than 6 g of glucose, which approximately corresponds to three cubes of sugar [120]. Since the blood is not a large reservoir of glucose, the large amount of glucose intake is added to the blood gradually by the digestive tract and the liver, and removed by the cells of the body for hours. Like all physiological systems (stated in Chapter 1), the underlying process of the glucose regulatory system is *nonlinear*, *stochastic*, and *complex*, all of which make an external control over the glucose concentration difficult.

Various mathematical models have been developed to describe glucose variations using physiologically-oriented models or data-driven approaches. The physiological modelling is mainly based on compartmental modelling, as it approximates a complex system by a number of interconnected subsystems named compartments. Compartmental models have been used to study the dynamic flow of chemicals such as nutrients, hormones, drugs and radioisotopes between different organs of the human body. In these cases, a ‘*compartment*’ is defined as an amount of a chemical that acts kinetically in a homogeneous way; if the chemicals in the same physiological space behave differently in their kinetics, they are considered to belong to different compartments [121]. The flows of chemicals between the compartments follow physical rules, which can be expressed as mathematical ordinary differential equations [121].

The first simple linear two-compartment model [122, 123] aimed at describing the blood glucose regulatory system during a glucose tolerance test, which monitored glucose concentration variations for more than three hours after a large dose of glucose intake [124]. Following this publication, the slightly more complex ‘minimal model’ [125] gained wide popularity and still remains the most popular choice for diagnostic purposes [126]. The minimal model separates the effect of insulin on glucose utilisation into two compartments: the glucose-dependent insulin compartment, and the postulated glucose-independent insulin compartment. The parameters of the minimal model (insulin sensitivity and glucose effectiveness) have been shown to have clinical importance and can be estimated from intravenous glucose tolerance test data [127] using nonlinear

least squares methods [128] or Bayesian techniques [129]. More complex models with more compartments [130] based on the minimal model [131–135] have been developed to capture the complexity of the underlying physiology (production, distribution, and degradation of glucose and insulin) by compartments, each of which associates with several differential equations. For example, Watson has built a three compartment model using a proportional-integral-derivative controller [134]. With a specially designed triple-tracer experiment, a ‘*Maximal Model*’ (MM) was proposed [135] including 12 differential equations and 35 parameters.

For all these compartmental models, the parameters, which have direct physiological interpretations, need to be obtained or inferred from measurements. Identifying these parameters is difficult, sometimes impossible, because the data collection for certain variables may be too expensive or unethical in practice. For example, all compartmental models require readily available time series for blood glucose and insulin concentrations [112]. While the glucose concentration time series are relatively easy to obtain, attaining enough insulin concentrations for an individual requires invasive and costly blood tests. In practice, clinicians and patients observe the glucose variations to improve DM management. Therefore, mathematical models that only require the glucose time series are in urgent need to give information about the underlying glucose regulatory system for the patient.

Data-driven models are capable of exploiting the information hidden in the data without detailed knowledge of the underlying physiological processes [136]. Modern continuous glucose monitoring (CGM) devices [137] – developed for the purpose of improving glucose self-monitoring management and avoiding dangerous hypoglycemic episodes – are able to obtain hundreds of data points collected at short time intervals, and have become particularly valuable for data-driven modelling. Many data-driven methods have been developed including artificial neural network models [120, 138], Volterra models [139] and others [112, 119]. Among them, one of the most popular data-driven methods is based on difference equations - autoregressive (AR) models. Sparacino et al. [140] suggested a first-order autoregressive model with time-varying parameters, and Gani [141] proposed an AR model of order 30 with fixed coefficients. AR models represent the random process of glucose concentration as a linear combination of past values and external inputs. However, to reflect the complex underlying process, a small number of AR model coefficients is not sufficient to capture the temporal variations of the time

series [142]; but an AR model with a large number of coefficients suffers from overfitting the data by mistakenly treating the noise as a feature in the time series.

On the other hand, differential equation based data-driven models have also been explored and showed promising results [108, 131, 143]. Ordinary differential equations (ODE) permit a more parsimonious presentation of data compared with a high order difference equation. Wu [108] modelled the blood glucose excursions after food intake by a second order linear differential equation based on CGM data from one T2D patient. Khovanova et al. [144] demonstrated that such linear systems can successfully describe some postprandial (after food intake) glucose excursions in subjects without DM, whereas strong nonlinear responses were evident for many DM profiles. It is concluded that the blood glucose variation in the diabetes profile is greater in amplitude and smoother, with retention of inter-dependence between neighbouring values in a profile. Even though the form of the nonlinearity is not explored, the importance of including *nonlinear* terms in the modelling equations is highlighted in the paper [144].

Another factor characterising glucose dynamics is the presence of a strong *stochastic* component [56, 131]. As stated in Chapter 1, physiological systems are intrinsically stochastic, which can be incorporated in the form of system noise into the model. System noise accounts for model misspecification, and can be added to an ODE system as a stochastic term to form ‘*stochastic differential equations*’ (SDE). By using models based on SDEs, the effects of model uncertainties can be decoupled from the effects of measurement noise. From the parameter estimation point of view, using a SDE instead of an ODE can decrease serially correlated residuals, high values of which are an indicator of model structure misspecification. However, SDE models have not been given the requisite attention in the vast literature on glucose regulation models.

It is worth noting that the solutions of ordinary differential equations can be presented in polynomial form. One may argue that polynomial fitting is more straightforward compared with ODE based models. However, an ODE based model has the advantage of structural interpretability compared to a model in a polynomial form. Even though this data-driven approach does not provide direct physiological interpretation for the model parameters, much information can be obtained from the inferred model structure and parameter values such as the order of the system, the class of the dynamic system, the inferred system and measurement noise intensity, and the natural frequency and

damping coefficient in the linear model. Another choice would be to develop models that are based on transfer functions. However, the main limitation of transfer functions is that they can only be used for linear system [145]. It should be pointed out that despite this limitation, transfer functions still remain a valuable tool for designing controllers for nonlinear systems, mainly through constructing their linear approximations around an equilibrium point of interest. However, models based on transfer functions cannot take system noise into account, which contradicts the nature of the underlying physiological system. On the other hand, an ODE based model has the flexibility of incorporate nonlinearity and stochasticity into the model framework.

For people with or in progression towards DM, it is important to maintain glycaemic stability during daily activities. A personalised model that accurately describes the glucose dynamics can help to monitor and improve an individual's control over DM in their daily lives. Therefore, the aim of this research is to develop a data-driven continuous-time model to describe the transient response of blood glucose dynamics to each food intake. The model is in the form of SDEs with a minimal number of equations and parameters, and accounts for nonlinearity and stochasticity of the underlying glucose dynamics. The VB method described in Chapter 2 is applied to select the best model structure among several model candidates and infer the deterministic and stochastic model parameters. Other methods, such as least squares methods, that could not serve the purpose of inferring the parameters and selecting the model based on stochastic differential equations are therefore not considered in the chapter. Sampling methods, such as Markov Chain Monte Carlo methods, are not the main focus of this research, and the advantages and disadvantages of these methods compared with VB methods have been discussed in detail in Section 2.2.2.

3.2 Data description

Glucose profiles from fifteen subjects, including five subjects without DM (control group), four subjects with T1D and six subjects with T2D, were collected by our clinical collaborator from the University of Oxford, Dr. Tim A. Holt, and were available for the study [144]. The recruitment ensured a diverse sample of ages and treatment regimens. Baseline biographical data were obtained on age, sex, body mass index, type of DM, treatment regimen, and recent HbA1c value which can be found in Table 3.1. The

treatment regimen listed in Table 3.1 is assumed to be consistent throughout the whole experimental period, and does not affect the modelling development in this study, because the main focus of this research is the dynamics of the glucose response following food intake. Medtronic Minimed CGM devices were used to obtain blood glucose values every 5 minutes over 72 hours, and one participant in the T1D group voluntarily (on the independent advice of their clinician) kept the device on for more than 6 days. The CGM devices use enzymatic sensors that are inserted subcutaneously in the abdomen to measure the interstitial fluid glucose concentration. There are 3 – 12 minutes of time lag between the interstitial fluid glucose concentration and the plasma glucose concentration due to the diffusion of glucose across the capillary endothelial barrier [137]. Measurement time series $G(t_i)$ are available for each subject, and comprise the glucose concentrations (in mmol/L) at time points $t_i = ih$, where $h = 5$ minutes is the sampling interval and $i = 1, 2, \dots, N$ (N is the number of measurement points). The measurements were taken in ‘free living’ conditions, i.e. no restrictions were placed on the subjects’ daily activities or food intake. The types of meals taken and the calorific intake are not available.

Several example time series are presented in Fig. 3.2. The dotted peaks in the time series represent the postprandial blood glucose excursions. To avoid mistaking measurement noise for genuine postprandial peaks, only distinguishable peaks with height more than 1.1 mmol/L during the daytime from 6 am to midnight were selected. The highest peak value for the subjects in the control group is just below 8 mmol/L, whereas the highest values for the T1D and T2D patients are greater than 15 mmol/L. Because the CGM devices need calibration at the beginning of each experiment and all the subjects started wearing the device during the evening time of the first day, the first peak for all the participants is selected as the first peak of the second day, which can be seen in Fig. 3.2. Since there are no restrictions on the time and the number of meals (food intake), the subjects have various numbers of distinguishable peaks corresponding to food intake events (between five to fourteen peaks for each individual).

There was no detailed information about the time of the food intake. The glucose concentration time series measured in the interstitial fluid by the CGM device was known to have a lag behind the blood glucose concentration. Therefore, the beginnings of the transient responses to each food intake needed to be determined by the modeller. After the transient response, the glucose concentration settled to a steady state. It was common in the fifteen time series that another external excitation (such as food

TABLE 3.1: Biometric indices, treatment regimens, HbA1c values and corresponding estimated average blood glucose levels of participants.

No.	Age	Sex	BMI <i>kg/m²</i>	Diabetes status	Treatment regimen	HbA1c mmol/mol	Glucose level mmol/L
1	57	F	20.5	T1D	Basal bolus (glargine plus aspart)	63	10.0
2	27	F	19.2	Control	N/A	N/A	N/A
3	59	F	27.3	Control	N/A	N/A	N/A
4	49	F	21.9	Control	N/A	N/A	N/A
5	32	F	29.4	T1D	Insulin pump	55	9.0
6	74	M	20.5	T2D	Metformin, gliclazide, rosiglitazone	61	9.7
7	66	F	25.9	T1D	Insulin pump	38	6.3
8	75	M	23.4	T2D	Metformin	46	7.6
9	68	F	32.7	T1D	Basal bolus (glargine plus aspart)	48	7.8
10	39	F	21.3	Control	N/A	N/A	N/A
11	61	F	32.6	T2D	Metformin	52	8.4
12	56	M	30.0	T2D	Metformin	68	10.8
13	52	F	44.5	T2D	Metformin, glargine	89	13.8
14	22	F	19.6	Control	N/A	N/A	N/A
15	63	F	27.0	T2D	Newly diagnosed	42	7.0

intake, exercise, etc.) occurred before the glucose concentration had the chance to settle. In these cases, the response to the first food intake was interrupted and the steady state, which was the basal level G_b , remained unknown. The basal glucose level usually demonstrates slow nonstationary dynamics [144]. Therefore, the basal level for each individual peak was considered different. Compared with the baseline differences from peak to peak, the baseline variations within the duration time of a single peak could be neglected and the baselines were considered as constant during the transient response to one food intake.

The shapes of the peaks exhibit different patterns, even for the same subject. The most

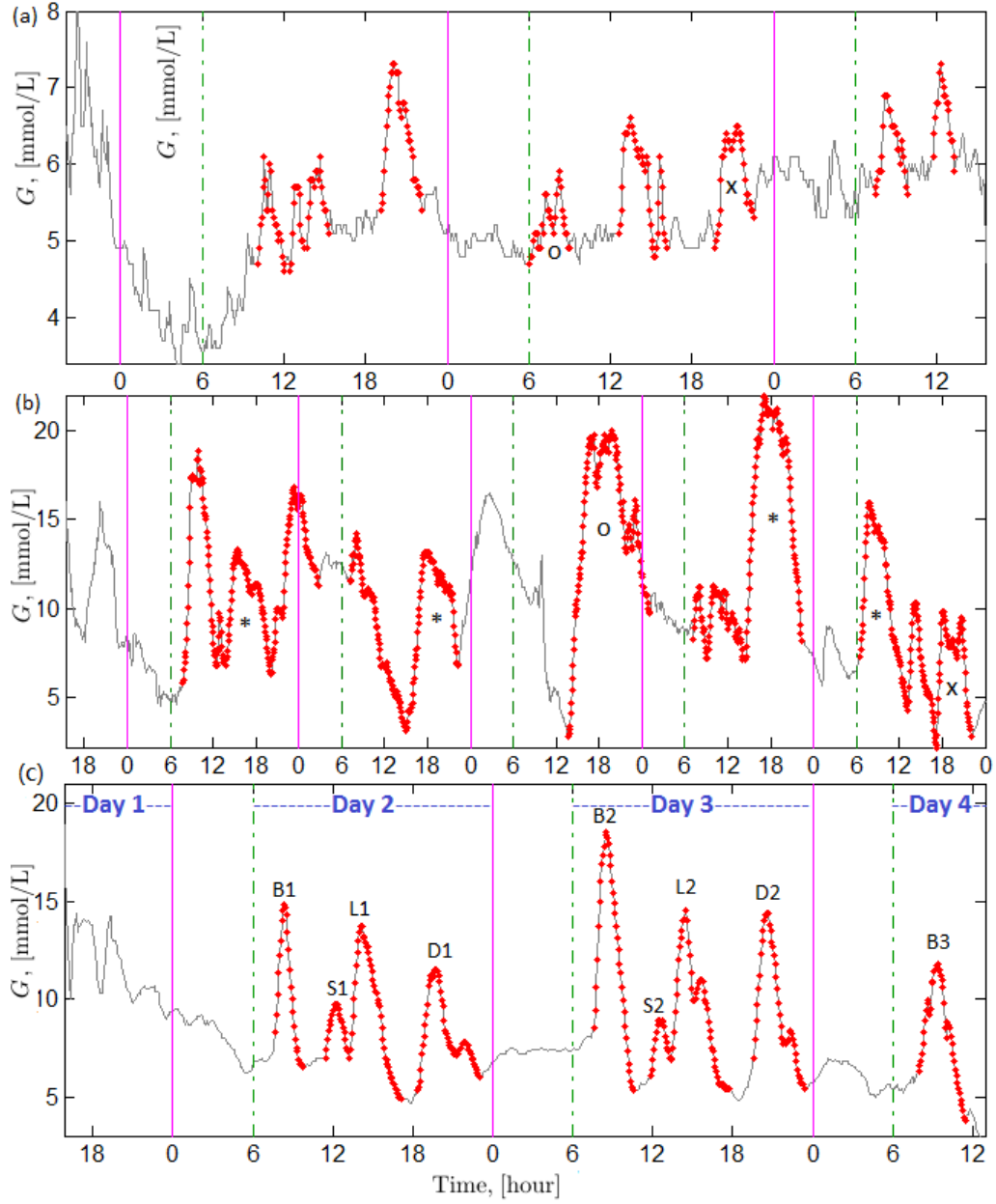


FIGURE 3.2: Example subcutaneous glucose time series G of a participant from (a) the control group (b) the T1D group (c) the T2D group. The solid grey curves represent the measured glucose values and the dots are the values used for modelling of single postprandial peaks. The solid and dashed vertical lines correspond to midnight (0 hours) and 6 am respectively. The first several hours of data in Day 1 (to the left from the first solid vertical lines) were excluded from the modelling due to the adjustment period of the CGM system. 'B' indicates breakfast, 'S' indicates snack, 'L' indicates lunch, 'D' indicates dinner.

common patterns are as follows: 1) peaks with a clear and almost symmetrical rise and fall, such as the first peak in Fig. 3.2 (c); 2) peaks with a steep rise followed by a slow fall and then a rapid fall, such as four peaks marked with '*' in Fig. 3.2 (b); 3) peaks with a distinguishable secondary peak indicated with 'x', or with multiple subpeaks

indicated with ‘o’ in Fig. 3.2 (a) and (b). According to [146], the glucose variations after food intake (and therefore the shape of the curves) can be influenced by different macro-nutrients of the meal, the time of the meal, the gender of the subjects and the diurnal cycling of hormones [146]. The peak shape variation is most likely due to erratic gastric emptying. The appearance of glucose in plasma depends on the gastric emptying rate. Liquids display exponential gastric emptying without an initial lag, while solids show biphasic gastric emptying with a lag [147]. When there are multiple peaks existing within a short period of time (within 3 hours), they may be caused by various emptying rate for different meal compositions or a pulsation secretion of insulin for the same food intake event, but can also be caused by two separate food intake events within a short interval of time. The complex patterns and the varieties in peak dynamics, as well as the lack of information on the precise food intake (including time, quantity and the constitution of macro-nutrients) make the task of finding a universal model difficult.

Observing the time series in Fig. 3.2 can provide valuable information about the subjects’ food routines. For example, the peaks representing breakfasts for the subject in Fig. 3.2 (c) are similar suggesting the subject has a daily breakfast routine. The small peaks between the breakfast and lunch in both days suggest that the subject also has a routine of eating a snack before lunch. Other subjects, such as the time series shown in Fig. 3.2 (b), have less fixed meal times and more varieties of features in peaks that represent the same meal in different days.

3.3 Model and Methods

3.3.1 Model formulation

Three important characteristics of the blood glucose response to food intake need to be taken into account during model formulation. First, the glucose response is nonlinear (as explained in Section 3.1). Second, stochasticity enters into the model in two forms: 1) as measurement noise from the CGM device, 2) as dynamic intrinsic noise accounting for model misspecification resulting from factors other than food intake, including physical activity and emotional stress. For patients using insulin, inaccurate estimation of the necessary dose is another factor influencing postprandial excursions. Third, the endocrine system tends to maintain homeostasis, and any deviations of blood glucose

from the basal level decay rapidly and return to the pre-disturbed status. As described in Section 3.2, the basal glucose level is different throughout the day and demonstrates slow nonstationary dynamics [144]. In this study the basal level G_b of glucose is assumed to be constant within one peak (usually several hours), but different from peak to peak. Compared with the minimal model where fixed baselines are considered, a model with a different basal level for different peaks gives a more realistic view of the slow variation in the basal level, which might be caused by the circadian rhythm. The basal level is chosen between the first data point when the transient response to the food intake starts, and the last data point when the glucose concentration reaches its steady state. The glucose concentration over the baseline G_b represents the system's transient response to food intake.

The generic model framework (2.2) constituting a system equation and a measurement equation has been introduced in Section 2.1.1. To achieve a parsimonious description of the postprandial excursions for all the peaks, a minimal order of the system and a minimal number of parameters are required. Wu [108] has developed a model based on a second order linear ODE that is capable of describing the postprandial blood glucose excursions for one T2D subject. Inspired by this publication, an order of two is selected as the system order. Therefore, the model proposed in Section 2.2 with an order or two was selected to describe the transient glucose response to food intake for people with and without DM namely:

$$\ddot{x}_t + f_1(x_t)\dot{x}_t + f_0(x_t) = u_t + \eta_t \quad (3.1a)$$

$$y_t = x_t + \varepsilon_t \quad (3.1b)$$

where (3.1a) is the system equation. The state vector of the system is $(x_t, \dot{x}_t)^\top$, η_t corresponds to the dynamic noise, u_t is the input function. Equation (3.1b) is the measurement equation, y_t is the measurement value at time t and ε_t corresponds to the measurement noise. Two stochastic terms, the system noise η_t and the measurement noise ε_t , were modelled as additive white Gaussian noise (refer to Section 2.1.2 for definition) [148] with zero means and intensities of I_η and I_ε respectively.

With no accurate information about how long each food intake lasts and how fast the food absorbed through the digestion tract can be reflected in the rise of the blood glucose level, the external excitation input u_t to the system is simplified as a bolus injection of

glucose at time zero. This means that the input function u_t equals zero when $t > 0$, and equals F when $t = 0$. Mathematically, such an input can be formulated as follows:

$$u_t = F\delta(t) = \begin{cases} F & \text{if } t = 0 \\ 0 & \text{if } t \neq 0 \end{cases} \quad (3.2)$$

where $\delta(t)$ is the Dirac delta function that is zero everywhere except at zero, with an integral of one over the entire time line [149]. F is regarded as an unknown food impact factor.

Functions $f_1(x_t)$ and $f_0(x_t)$ play an important role in the suggested model. Shown by Khovanova et al. [144], nonlinearity needs to be introduced into the model to have the capacity to describe the dynamic behaviours in all of the time series. As introduced in Section 2.1.1, to describe a variety of dynamic responses to the external excitation force (representing the food intake), the functions $f_1(x_t)$ and $f_0(x_t)$ in the system equation need to be structurally flexible. Polynomial forms allow the description of a wide range of nonlinear solutions, and can be adjusted to fit different dynamic features by increasing the number of components and by varying their parameters. Therefore, $f_1(x_t)$ and $f_0(x_t)$ are defined as follows:

$$f_1(x_t) = \sum_{i=0}^n \theta_{k_i} x_t^i, \quad f_0(x_t) = \sum_{j=1}^m \theta_j x_t^j \quad (3.3)$$

where θ_{k_i} and θ_j in (3.3) are system parameters.

A linear deterministic equation is a particular case of this model when $n = 0$, $m = 1$, $f_1(x_t) = \theta_{k_0}$ (later referred to as θ_k in linear case for short), $f_0(x_t) = \theta_1 x_t$ and the stochastic terms are zero: $\eta_t = 0$, $\varepsilon_t = 0$. The linear noise-free system has been considered in [108] and was based on one T2D patient profile. The nonlinearity is introduced into the system by including the higher order polynomial functions in (3.3).

3.3.2 Model selection and parameter inference

Model (3.1a – 3.1b) covers a variety of dynamic patterns depending on the order of the polynomial functions $f_1(x_t)$ and $f_0(x_t)$ (later referred to as f_1 and f_0 for short). The aim was to select the best – parsimonious but not over-simplified – model structure for

each postprandial peak with the *minimal* number of parameters. To introduce sufficient nonlinearity into the system without overcomplicating the model, combinations of f_1 and f_0 with different numbers of polynomial terms have been tried to fit the time series. For example, three polynomial terms were used in both functions f_1 and f_0 for the first model candidate M_1 as listed in Table 3.2. The second model candidate M_2 has three terms in f_1 and one linear term in f_0 . The third model candidate M_3 has three terms in f_0 and one term in f_1 . All three models have been compared with each other and with the linear model M_L . The combinations described by models M_1 , M_2 , M_3 and M_L are not exhaustive. Many combinations were disregarded before applying the model selection and Bayesian inference method. For example, omitting the quadratic term of f_0 in model M_3 would make the system highly symmetrical which would significantly restrict the possible forms of the solutions. Similarly, omitting the cubic components in f_0 would lead to unstable solutions of the system, which can be proven analytically [150]. Also, using four or more terms would over-complicate the structure of the system, which is against the aim of searching for a parsimonious model.

TABLE 3.2: Four model candidates for fitting

Models	f_1	f_0
M_1	$\theta_{k_0} + \theta_{k_1}x_t + \theta_{k_2}x_t^2$	$\theta_1x_t + \theta_2x_t^2 + \theta_3x_t^3$
M_2	$\theta_{k_0} + \theta_{k_1}x_t + \theta_{k_2}x_t^2$	θ_1x_t
M_3	θ_k	$\theta_1x_t + \theta_2x_t^2 + \theta_3x_t^3$
M_L	θ_k	θ_1x_t

To decide which of the four models best describes the postprandial transient blood glucose response and, at the same time, identify corresponding values of the deterministic (θ_{k_i}, θ_j) and stochastic (I_η, I_ε) parameters, an inference method that can incorporate nonlinearity and stochasticity into the model framework was required. Both sampling methods such as the Monte Carlo methods (refer to Section 2.2.2 for detail) or the VB method can serve this purpose well. The benefit of using the VB method compared to the Monte Carlo methods has been elaborated in Chapter 1 and Section 2.2.2, and therefore the VB method was employed. The algorithm accounts for the stochastic nature of the underlying glucose dynamics and also enables us to distinguish between these two types of stochasticity: dynamic and measurement noise. The algorithm is flexible and estimates the model parameters in the form of probability distributions rather than fixed

values. The final result thereby includes information on uncertainties in the parameter estimates. Specifically, the VB toolbox [77] was applied. The equations (3.1a) – (3.1b) were incorporated into the framework of the VB algorithm. Free energy values \mathcal{F} for each model candidate M (where M is M_1 , M_2 , M_3 or M_L) were found, and the model with the highest value of free energy (details in Section 2.3.1), if it also satisfied other criteria explained in Section 3.3.3, was selected.

As elaborated in Section 2.5 and Section 2.8.2, the choice of priors and hyperpriors is important for the VB method since it influences the posterior parameter distributions. Since there was no information about the parameter values in the model, weakly informative priors were selected for all deterministic parameters. Since the deterministic parameters were modelled as Gaussian distributions, a relatively large variance is important for the VB algorithms to search over a wide enough parameter space to look for the optimal values of the parameters. The mean values of the priors for the system parameters θ_k and θ_1 were set to zero in accordance with Wu’s paper [108]. For consistency, mean values of priors for extra deterministic parameters (θ_{k_1} , θ_{k_2} , θ_2 , θ_3) were also set to zero for models $M_1 - M_3$. All of the variances for the deterministic parameters were set to be 10^4 as a default value. To make sure that this variance was set sufficiently large, larger variances were used to see if the posterior distribution of the parameters would change. The differences are negligible, and therefore variances were selected as 10^4 .

The prior of the food impact factor F was set to be normally distributed with a mean value of $y_{t_2} - y_{t_1}$, where y_{t_1} and y_{t_2} are the first and second time points of each peak, with a variance of 10^4 . The noise precisions α_ε and α_η , which are inversely proportional to the noise intensities I_ε and I_η , were modelled by Gamma distributions with both shape a and rate b parameters set to 0.001 as the default hyperpriors, which was justified in Section 2.5. As illustrated in the examples shown in Section 2.8.2, the inference result can be sensitive to the choice of the prior and hyperpriors, and therefore different prior and hyperprior combinations for the different models were trialled to optimise the free energy for each model candidate. The models that are finally selected are proven locally identifiable (see details in Sections 3.4.2).

3.3.3 Model selection criteria

Four models (M_L , M_1 , M_2 and M_3) were compared to achieve the most satisfactory fitting. To select the best model, the following two criteria were considered:

1. Decay ratio. Since the glucose dynamics are tightly regulated by the endocrine system, it is a stable process with limited decaying oscillations except for patients in a critical condition or T1D patients who have to maintain the blood glucose concentrations with insulin injections. Thus, any blood glucose excursion after food intake should relax to a steady state demonstrating the dynamics of an over-damped system. In the available time series the transient response corresponding to one food intake is often superimposed by the response of another food intake within a short interval, and such overlapping responses cause problems of observing and identifying some important dynamic features, such as the oscillation rate. We believe, however, that any model with a decay ratio (ratio of the height of the second peak to the height of the first peak) higher than 33 % is unrealistic, and should be disregarded.

2. The difference in the free energy \mathcal{F} between any two models: $|\Delta\mathcal{F}|$. According to [80], if $|\Delta\mathcal{F}| > 3$ (3 is a conventional choice based on [81] and [80], refer to Section 2.4 for further detail) in Chapter 2), there is strong evidence that the model with the higher value of the free energy is better. This rule is derived from a well known Bayes' factor which is a measure for comparing the evidences of two Bayesian models [80].

3.4 Results and Discussions

3.4.1 Model selection and parameter inference

All the distinguishable peaks with height more than 1.1 mmol/L starting between 6am to midnight were selected, and all selected 132 time series comprising blood glucose dynamics after food intake were fitted using the models M_1 , M_2 , M_3 and M_L . Note that it is an event-based model so that each time series represents one food intake event. Fig. 3.3 shows (typical) fitting for one postprandial peak. The procedure of model selection is illustrated by considering this typical example. The most complex nonlinear model M_1 produced an unstable result with unrealistic periodic oscillations over time, leading to a decay ratio close to one that does not satisfy the second criterion in Section

3.3.3. Therefore, M_1 was disregarded. M_3 results in an unstable solution, leaving M_L and M_2 for further consideration.

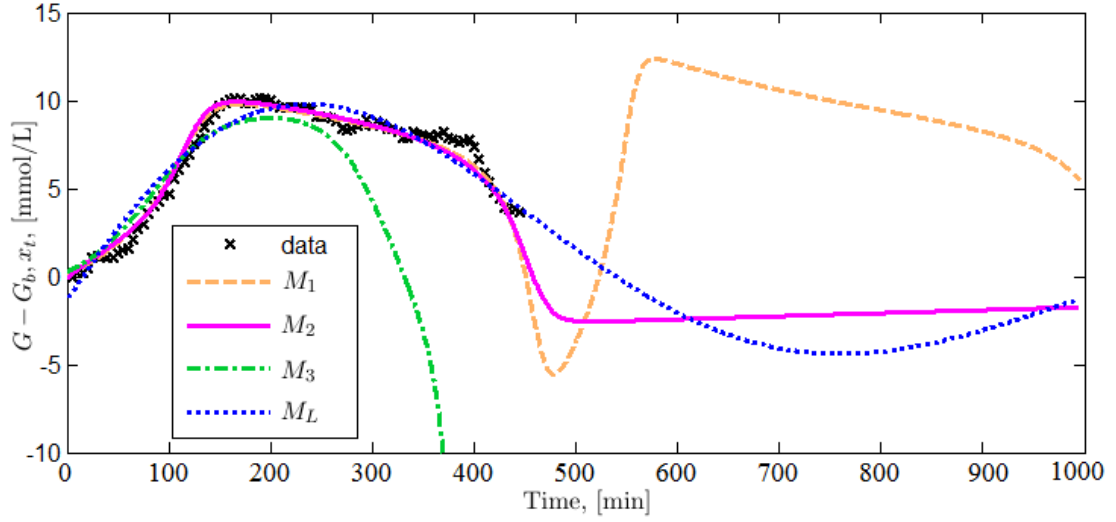


FIGURE 3.3: Typical outcome for one peak (the sixth peak in Fig. 3.2 (b)) fitting by four models.

The selection between the models M_L and M_2 can be influenced by the choice of the hyperpriors (a_1 and b_1) for the system noise. It is worth noting that the fitting results of M_1 and M_3 can also be influenced by the choice of the hyperpriors. But in general, M_1 and M_3 tend to have less stable solutions compared with M_L and M_2 , as high degree polynomials may suffer from instability. The goal of this research is to find a generalised model that can describe the glucose dynamics of food intake so that comparisons can be made between different meals within the same subject, between different subjects within the same group, and between different groups. Therefore, M_1 and M_3 will be not be further discussed. As seen in the example presented in Section 2.8.2, different settings of a_1 and b_1 cause large differences in the free energy value. Selecting the best model based on the default prior of $a_1 = b_1 = 10^{-3}$ (which is asymptotically the Jeffreys prior as shown in Section 2.5) may lead to a wrong choice of model. Table 3.3 and 3.4 present the values of the free energy using different hyperprior combinations for M_L and M_2 respectively to fit the time series shown in Fig. 3.3.

Considering M_L and M_2 were competing models fitted to the same data, same hyperpriors should be applied to them, i.e. using the values from the same location in Table 3.3 and 3.4. When hyperpriors a_1 and b_1 were both 10^{-3} , the free energy \mathcal{F}_{M_L} was higher (the values instead of the absolute values are compared) than \mathcal{F}_{M_2} indicating that M_L is the better model; instead, when a_1 was 10^{-1} and b_1 was 10^{-3} , \mathcal{F}_{M_2} was larger than

TABLE 3.3: Free energy of M_L , denoted as \mathcal{F}_{M_L} , for different hyperprior settings of the precision of system noise

\mathcal{F}_{M_L}	$b_1 = 10^{-3}$	$b_1 = 10^{-2}$	$b_1 = 10^{-1}$	$b_1 = 10^0$	$b_1 = 10^1$
$a_1 = 10^{-3}$	-93	-95	-767	-82	-248
$a_1 = 10^{-2}$	-95	-71	-273	-72	-875
$a_1 = 10^{-1}$	-64	-152	-806	-83	-318
$a_1 = 10^0$	-106	-66	-991	-74	-239
$a_1 = 10^1$	-68	-85	-76	-69	-214

TABLE 3.4: Free energy of M_2 , denoted as \mathcal{F}_{M_2} , for different hyperprior settings of the precision of system noise

\mathcal{F}_{M_2}	$b_1 = 10^{-3}$	$b_1 = 10^{-2}$	$b_1 = 10^{-1}$	$b_1 = 10^0$	$b_1 = 10^1$
$a_1 = 10^{-3}$	-111	-78	-190	-57	-158
$a_1 = 10^{-2}$	-39	-111	-86	-190	-118
$a_1 = 10^{-1}$	15	-39	-111	-93	-196
$a_1 = 10^0$	16	15	-45	-112	-116
$a_1 = 10^1$	-40	-27	17	-88	-52

\mathcal{F}_{M_L} , indicating that M_2 is the better model. Thus, the selection of the model is not independent on the selection of hyperpriors. It is advisable to optimise the hyperpriors to obtain the largest possible free energy value for each model (M_L and M_2) first before choosing a model. As shown in Table 3.3 and 3.4, the largest three free energy values obtained using different combinations of a_1 and b_1 for each model have been marked **bold** in the table. Among them, only two of them shared the same location in both tables, indicated by circles around the free energy values. Between them, $a_1 = 10^{-1}$, $b_1 = 10^{-3}$ has higher free energy values in M_L , and therefore were finally selected as the optimised hyperpriors for both models.

The inference results with these selected hyperpriors for both models M_L and M_2 are presented in Table 3.5. It is worth noticing that the mean values of the inferred system noise intensity and measurement noise intensity are smaller in M_2 compared with the values in M_L , and the standard deviations for both noise intensity are also smaller in M_2 . Smaller noise intensity implies that the M_2 fits the data better, which can be confirmed by a larger free energy value of $\mathcal{F}_{M_2} = 15$ for M_2 compared with the free energy value $\mathcal{F}_{M_L} = -64$ for M_L .

TABLE 3.5: Inference results of the parameters for the example shown in Fig. 3.3

		Mean \pm SD	Unit
M_L	θ_k	$(3.20 \pm 0.06) \times 10^{-3}$	$[\text{min}^{-1}]$
	θ_1	$(3.82 \pm 0.11) \times 10^{-5}$	$[\text{min}^{-2}]$
	x_0	(-1.49 ± 0.03)	$[\text{mmol/L}]$
	\dot{x}_0	$(8.40 \pm 0.42) \times 10^{-2}$	$[\text{min}^{-1} \text{ mmol/L}]$
	I_η	$(4.03 \pm 0.30) \times 10^{-5}$	$[\text{min}^{-2} \text{ mmol}^2/\text{L}^2]$
	I_ε	(0.26 ± 0.019)	$[\text{mmol}^2/\text{L}^2]$
M_2	θ_{k_0}	$(1.80 \pm 0.75) \times 10^{-2}$	$[\text{min}^{-1}]$
	θ_{k_1}	$(2.09 \pm 0.34) \times 10^{-2}$	$[\text{min}^{-1} \text{ L/mmol}]$
	θ_{k_2}	$(2.60 \pm 0.33) \times 10^{-3}$	$[\text{min}^{-1} \text{ L}^2/\text{mmol}^2]$
	θ_1	$(6.40 \pm 1.12) \times 10^{-5}$	$[\text{min}^{-2}]$
	x_0	(-0.44 ± 0.14)	$[\text{mmol/L}]$
	\dot{x}_0	$(6.1 \pm 1.25) \times 10^{-2}$	$[\text{min}^{-1} \text{ mmol/L}]$
	I_η	$(2.50 \pm 0.19) \times 10^{-5}$	$[\text{min}^{-2} \text{ mmol}^2/\text{L}^2]$
	I_ε	$(6.3 \pm 0.48) \times 10^{-2}$	$[\text{mmol}^2/\text{L}^2]$

The same hyperpriors $a_1 = 10^{-1}$, $b_1 = 10^{-3}$ for system noise were applied to fit all the peaks in the cohort using M_L and M_2 . Some more examples illustrating how the choice between models M_L and M_2 was made are presented in Fig. 3.4. For the time series presented in Fig. 3.4 (a), the free energy difference between M_L and M_2 is 277.9 ($\mathcal{F}_{M_2} - \mathcal{F}_{M_L} = 111.3 - (-166.6) = 277.9$) indicating a strong nonlinear character in this peak. It is worth noting that the calculation of free energy does not take into account the simulated data points beyond the data period, and the extended simulation is to show the trend of the solution. The nonlinear model M_2 is able to capture the dynamics of this response better than the linear model M_L , even though M_2 still misses the sub-peak which has been described in Section 3.2. Such double (and multiple) peaks were commonly present in all the profiles. They can be explained by a biphasic gastric emptying rate [151] which has a big influence over the absorption rate of glucose [152]. Additionally, the release of insulin from the pancreas is not continuous [153] and consists of two stages, fast and slow, which may also cause the second rise in the postprandial excursion. For the DM subjects the peaks are generally higher and occur over a longer time period compared with those from the control group, and as such the effect of the sub-peak dynamics is more distinctive and increases the need for a nonlinear model to capture such features. For the time series presented in Fig. 3.4 (b), M_L does not satisfy

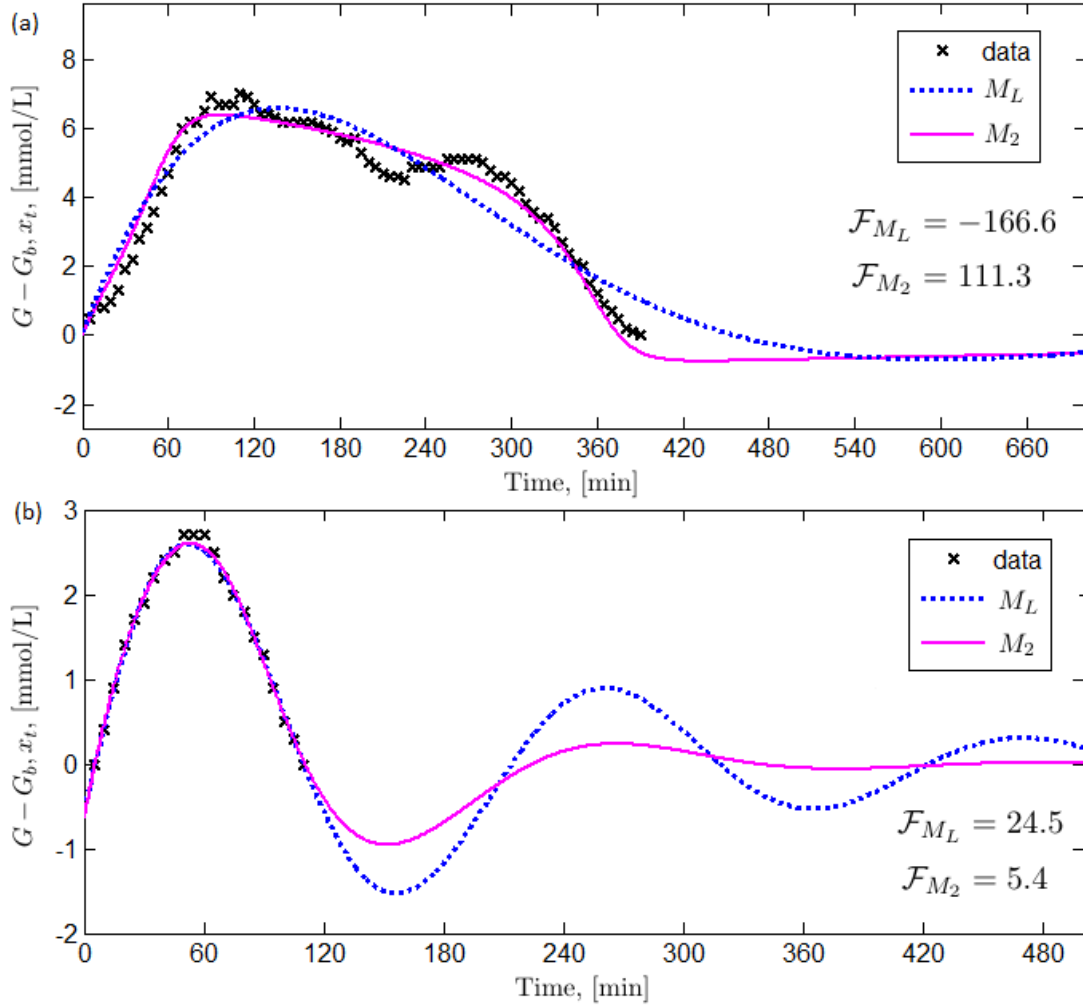


FIGURE 3.4: The fitting results are shown for: (a) a peak of a T2D profile; (b) a peak of a T1D profile. The lines are simulated deterministic solutions using the inferred parameters for M_L and M_2 .

the first criterion: the decay ratio is 35%. This means that the oscillations are considered to be unrealistic, and therefore M_2 was chosen as the fitting model.

Among 132 peaks across all fifteen subjects, the linear model M_L was selected for 56% of the peaks (Table 3.6) and the nonlinear model M_2 was selected for the rest of the peaks (44%). For the control group, M_L worked for 70% of the peaks; for the T1D and T2D group, only 54% and 45% of the peaks could be described adequately by M_L . More nonlinear peaks in DM groups than in controls emphasises the nonlinear character of response in people with DM. According to [154], the hyper- and hypo-glycaemic regions show greater deviation from linear behaviour, explaining the improved nonlinear fitting for the T2D group. M_L and M_2 are essentially the same model since M_L is a particular case of M_2 . They represent different levels of complexity. The reason that they have to

be treated as two models in practice is because the VB method treat parameter values as distributions, which means that even if the mean values of the two extra parameters in M_2 are inferred as zero, the VB methods calculates the free energy taking account of the uncertainty of these two parameters, leading to a smaller value of free energy for M_2 .

TABLE 3.6: Summary of peak fitting using M_L and M_2

Summary	Total peaks	Control	T1D	T2D
Total peaks	132	46	37	49
No.(%) fitted by M_L	74	32 (70%)	20 (54%)	22 (45%)
No.(%) fitted by M_2	58	14 (30%)	17 (46%)	27 (55%)

As mentioned in the last paragraph of Section 3.2, some subjects exhibit similar patterns in the glucose excursion after certain meals (breakfast/lunch/dinner) everyday. The pattern in breakfast is easier to identify due to the fact that breakfast is taken after an overnight fasting period and many people eat similar food for breakfast everyday. To investigate whether the breakfasts for the same patient can be described by the model with the same level of complexity, three peaks, which correspond to the glucose response of a T2D subject after breakfast for three consecutive days (see Fig. 3.5), have been fitted by M_L and M_2 as demonstrated in Fig. 3.6 (a) – (c). Even though there are no striking differences visually between the deterministic solutions of M_L and M_2 shown in Fig. 3.6 (b) – (c), the large differences in free energy values are caused by different uncertainties of the inferred parameters, the states and the hyperparameters. Please refer to Section 2.3.4 for more detail.

For each of the three peaks, visually there is minimal difference between the deterministic solutions from the inference results of the two models, but the free energies \mathcal{F}_{M_2} for M_2 in all these cases are larger than the free energies \mathcal{F}_{M_L} for M_L (as shown in Fig. 3.6 (a) – (c)). In all cases, the decay ratio was below 33% for both models, thus satisfying the first criterion in Section 3.3.3. Therefore, M_2 was chosen for all three peaks based on the second criterion in Section 3.3.3.

Out of the eleven subjects who ate breakfast regularly, either M_L or M_2 could be fitted to all breakfast peaks for seven of the subjects; for the remaining four subjects, some breakfast peaks selected M_L and the rest of the breakfast peaks selected M_2 . For

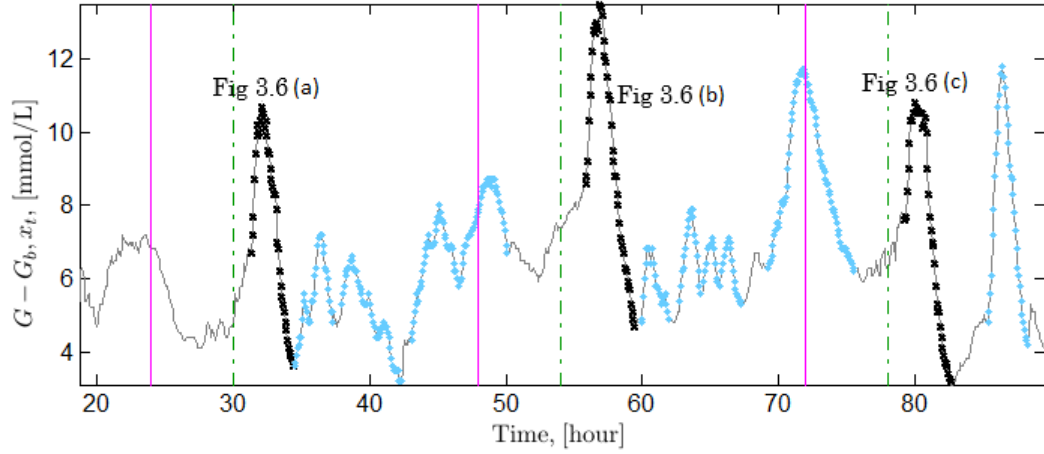


FIGURE 3.5: Glucose time series G of a T2D subject. The solid grey curves represent the measured glucose values and the dots are the values used for modelling of single postprandial peaks. The glucose time series corresponding to responses to food intake during breakfasts are indicated by dark black crosses. The solid and dashed vertical lines correspond to midnight (12 am) and 6 am respectively.

example, for the T2D subject shown in Fig. 3.2 (c), M_L was chosen for all the breakfast and dinner peaks (5 peaks), and M_2 was chosen for all the lunch and snack peaks (4 peaks). The selection of model between M_L and M_2 was performed strictly based on the criteria provided in Section 3.3.3. It is not surprising that different peaks belonging to one person can be fit by two models (M_L and M_2). M_L is a particular case of M_2 when θ_{k_1} and θ_{k_2} equal to zero. Computationally, when the parameters θ_{k_1} and θ_{k_2} are close to zero, the free energy of M_2 will be smaller than the free energy of M_L due to the penalisation for a larger degree of freedom in the parameter space. A different preference of the models for different meals implies different mechanisms that control the postprandial glucose excursions at different times of the day. The different mechanisms may be caused by diurnal hormone fluctuation and/or different compositions of meals at different times.

3.4.2 Structural identifiability and parameter sensitivity

A structural identifiability analysis has been performed for both models M_L and M_2 . As shown in the example with the system (2.68), it has already been proven that the linear model M_L is structurally identifiable. The nonlinear system M_2 , which is the example shown in Section 2.6, has also been proven identifiable.

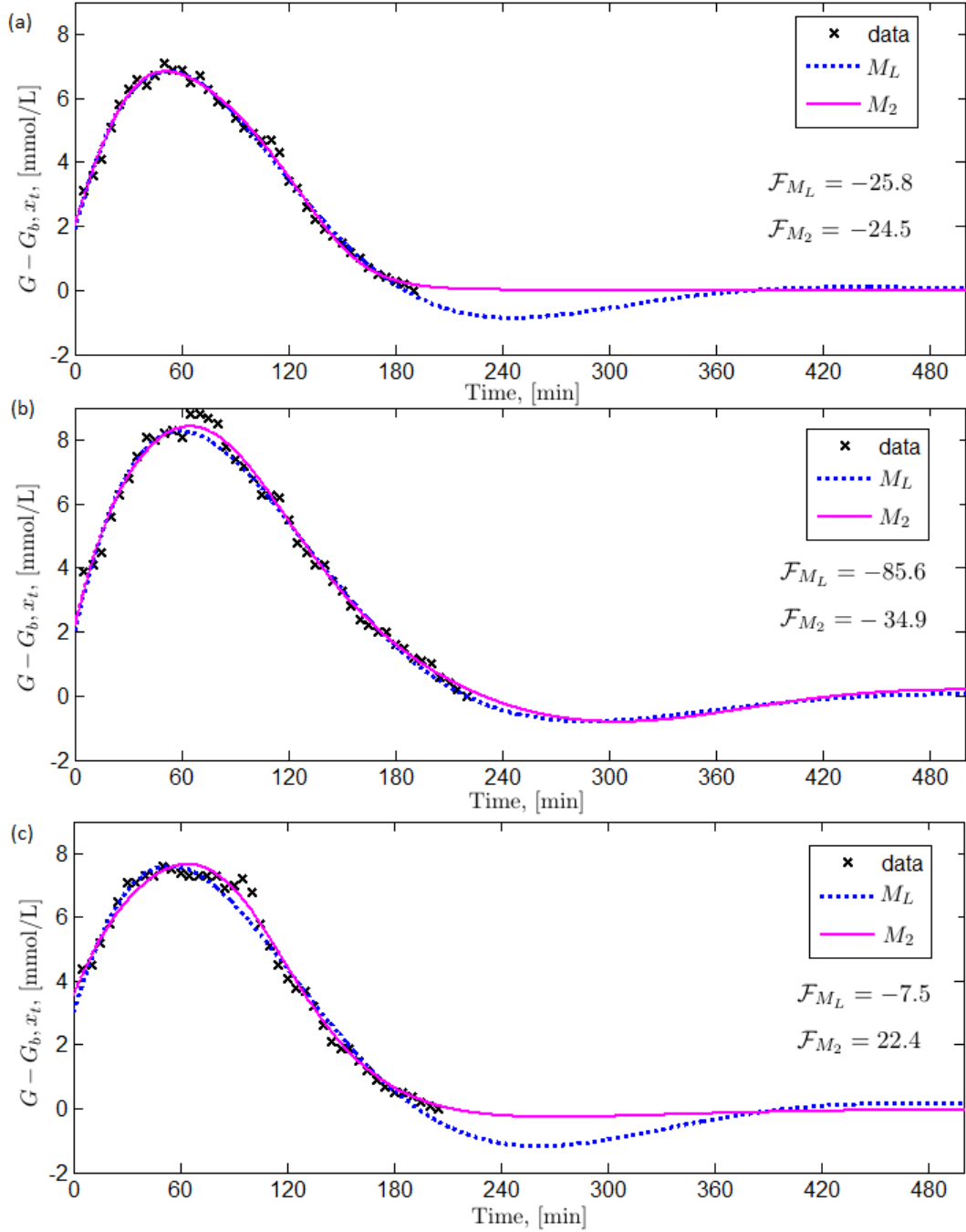


FIGURE 3.6: The fitting results are shown for three peaks (represent breakfasts) of a T2D profile (as in Fig. 3.5) in three consecutive days from (a) – (c). The free energy value of M_L is denoted as \mathcal{F}_{M_L} , and the free energy value of M_2 is denoted as \mathcal{F}_{M_2} .

To check parameter sensitivity, the procedures explained in Section 2.7. The one-at-a-time parameter sensitivity analysis has been performed for both models M_L and M_2 (see details in Section 2.7). To check how sensitive the output is to a small change in each deterministic parameter, the following steps have been performed for each model:

- 1) Simulate the time series with no measurement or system noise and obtain the root

mean square error (RMSE) between the noise free time series and the measurement time series.

- 2) Take 1000 random samples (1000 samples are considered large enough to make statistical analysis) of $\theta_i^{(j)}$ ($j = 1, 2, \dots, 1000$) from a uniform distribution from $0.99\hat{\theta}_i$ to $1.01\hat{\theta}_i$, where $\hat{\theta}_i$ is the posterior mean of θ_i .
- 3) Calculate the $\text{RMSE}^{(j)}$ values between the measurement values and each of the 1000 generated time series, using the sampled parameter $\theta_i^{(j)}$ and the posterior means of the rest of the parameters.
- 4) Using (2.83), obtain the sensitivity index SI for parameter θ_i .

The same procedure can be performed for the inferred parameters of all of the time series. Considering that the parameter values for the different time series fitted by M_L or M_2 are in the same neighbourhood, the sensitivity index SI for one typical time series as shown in crosses in Fig. 3.3 is presented as an example here. For the linear model M_L and the nonlinear model M_2 , as shown in Table 3.7 and Table 3.8, one percentage change in the parameter can cause the corresponding SI percentage change in RMSE. The RMSE values between the measurements and the output, generated when the posterior means of the parameters were used, is 0.98 mmol/L for M_L and 0.42 mmol/L. With 1% perturbation in each parameter shown in the first column of Table 3.7 and Table 3.8, the range of $\text{RMSE}^{(j)}$ is shown in the third column. The RMSE values remains within a small range; therefore, the models M_L and M_2 are robust around the inferred posterior means of the parameters.

TABLE 3.7: Summary of the parameter sensitivities for M_L and the range of RMSE with 1% parameter perturbation

Parameter	SI	$\text{RMSE}^{(j)}$ range
θ_k	2.15	0.96 – 1.00 mmol/L
θ_1	1.29	0.97 – 0.99 mmol/L

3.4.3 Parameter comparison between the groups

The deterministic part of M_L is represented by two parameters θ_k and θ_1 , and of M_2 by four parameters θ_{k_0} , θ_{k_1} , θ_{k_2} and θ_1 . The stochastic part is represented by I_η and I_ε

TABLE 3.8: Summary of the parameter sensitivities for M_2 and the range of RMSE with 1% parameter perturbation

Parameter	SI	RMSE ^(j) range
θ_{k_0}	-3.77	0.40 – 0.43 mmol/L
θ_{k_1}	5.61	0.41 – 0.49 mmol/L
θ_{k_2}	-2.58	0.41 – 0.45 mmol/L
θ_1	-0.18	0.416 – 0.423 mmol/L

in both models. Statistical analysis of the models' parameters was performed by using the Wilcoxon rank sum test. The null hypothesis of no difference between the groups of interest was tested at the 5% level of significance, and this is presented by p -values.

3.4.3.1 Coefficients of function f_0 and undamped frequency

The functions f_0 have the same structure for both the linear M_L and nonlinear M_2 models (Table 3.2), and contain the coefficient θ_1 . The square root of this parameter ($\sqrt{\theta_1}$) defines an undamped natural frequency in the system, at which a system would oscillate in the absence of any driving or damping force. The values of $\sqrt{\theta_1}$ were compared among the three groups (control, T1D and T2D) for models M_L and M_2 (Fig. 3.7 (a)). If there is a significant difference in the medians between any two groups, the p -values are marked by a star. For the peaks that were fitted by M_L , the median value of $\sqrt{\theta_1}$ in the control group is significantly higher than the median values in both DM groups, whereas there is no statistically significant difference between the T1D and T2D groups. For the peaks fitted by M_2 , the same result is observed (Fig. 3.7 (b)): $\sqrt{\theta_1}$ differs significantly between the control group and both DM groups. Thus, by analysing the undamped frequency of the models developed it is possible to distinguish between the cases with and without DM.

3.4.3.2 Coefficients of function f_1 and damping

The functions f_1 characterise damping in the systems and are represented by polynomials of different orders for M_L and M_2 . The nonlinear model M_2 contains a quadratic nonlinear 'damping' function and is characterised by three parameters, whereas the linear model M_L contains one damping coefficient parameters. These inferred coefficients have

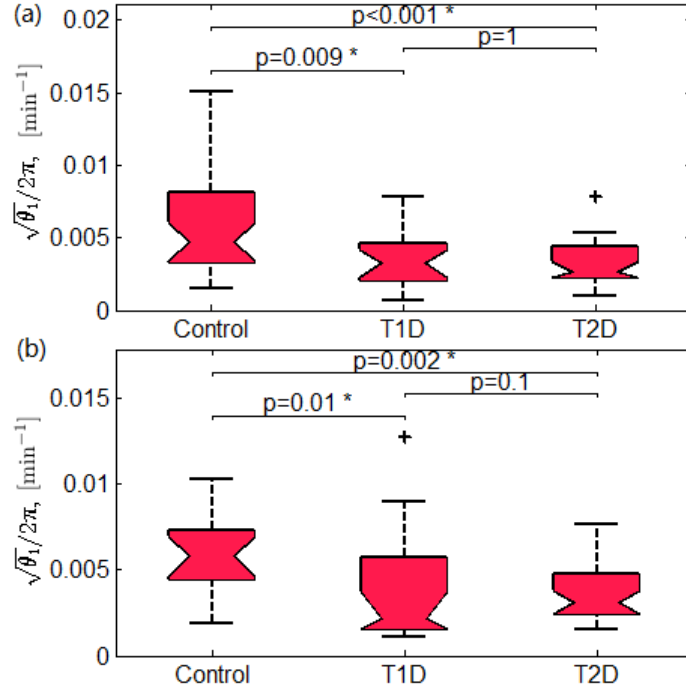


FIGURE 3.7: Boxplots for parameter $\sqrt{\theta_1}/2\pi$ obtained from (a) M_L and (b) M_2 . Note that the denominator 2π is to convert the units from $[\text{radian}/\text{min}]$ to $[\text{min}^{-1}]$

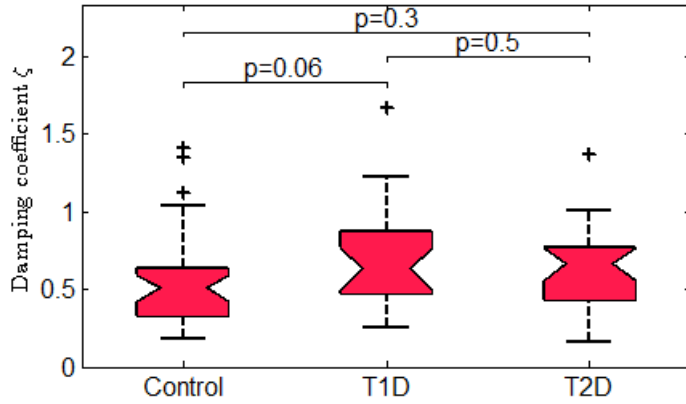


FIGURE 3.8: Boxplots for the damping coefficients $\zeta = \frac{\theta_k}{2 \times \sqrt{\theta_1}}$ obtained from M_L .

also been compared across the control and both DM groups. For the linear model M_L , the damping coefficient $\zeta = \theta_k/2\sqrt{\theta_1}$ and a boxplot for the three groups is shown in Fig. 3.8. No significant differences are shown between the three groups.

A comparison of the coefficients of the function f_1 between the three groups for both models M_L and M_2 is presented in Fig. 3.9. For M_L , the median value of θ_k of the control group is significantly larger than in the T2D group. For M_2 , it is noticeable from Fig. 3.9 (b) – (d) that the values of θ_{k_0} , θ_{k_1} and θ_{k_2} for the control group have a wider spread compared with the other two groups. The median value of θ_{k_0} for the control group is

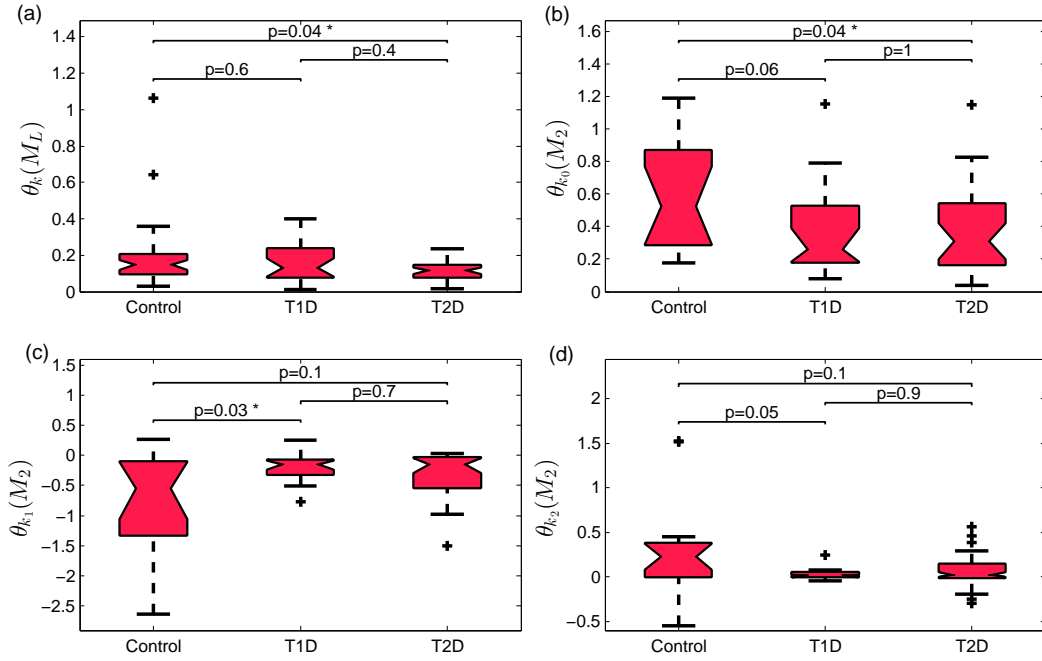


FIGURE 3.9: Boxplots for parameters: (a) θ_k in M_L ; (b) θ_{k_0} in M_2 ; (c) θ_{k_1} in M_2 ; (d) θ_{k_2} in M_2 .

significantly larger than the median value for the T2D group. The median value of θ_{k_1} for the control group is significantly smaller compared with the median value of the T2D group. However, there is no conclusive difference between the parameters θ_{k_2} , θ_{k_1} and θ_{k_0} of M_2 between the control group and both DM groups as the p -values are close to the threshold value of 0.05.

It is important to note that the parameter θ_{k_0} of the function f_1 defines the stability of both models, M_L and M_2 . In system theory, a system is in a steady state if the state variables which define the behaviour of the system are unchanging in time. It is assumed that the system would eventually reach a steady state with a constant value. Therefore, the real parts of the eigenvalues $\lambda_{1,2}$ of the linear system M_L or the linearised system M_2 around the steady state $[\frac{F}{\theta_1}, 0]^T$ must be negative:

$$\lambda_{1,2} = \begin{cases} -\frac{\theta_{k_0}}{2} \pm \frac{1}{2}\sqrt{\theta_{k_0}^2 - 4\theta_{k_1}} & \text{if } \theta_{k_0}^2 \geq 4\theta_{k_1} \\ -\frac{\theta_{k_0}}{2} \pm \frac{i}{2}\sqrt{-\theta_{k_0}^2 + 4\theta_{k_1}} & \text{if } \theta_{k_0}^2 < 4\theta_{k_1} \end{cases} \quad (3.4)$$

All the inferred parameters θ_{k_0} were positive as shown in Fig. 3.9 (b), making the real part of the eigenvalues negative and confirming the stability of the developed models.

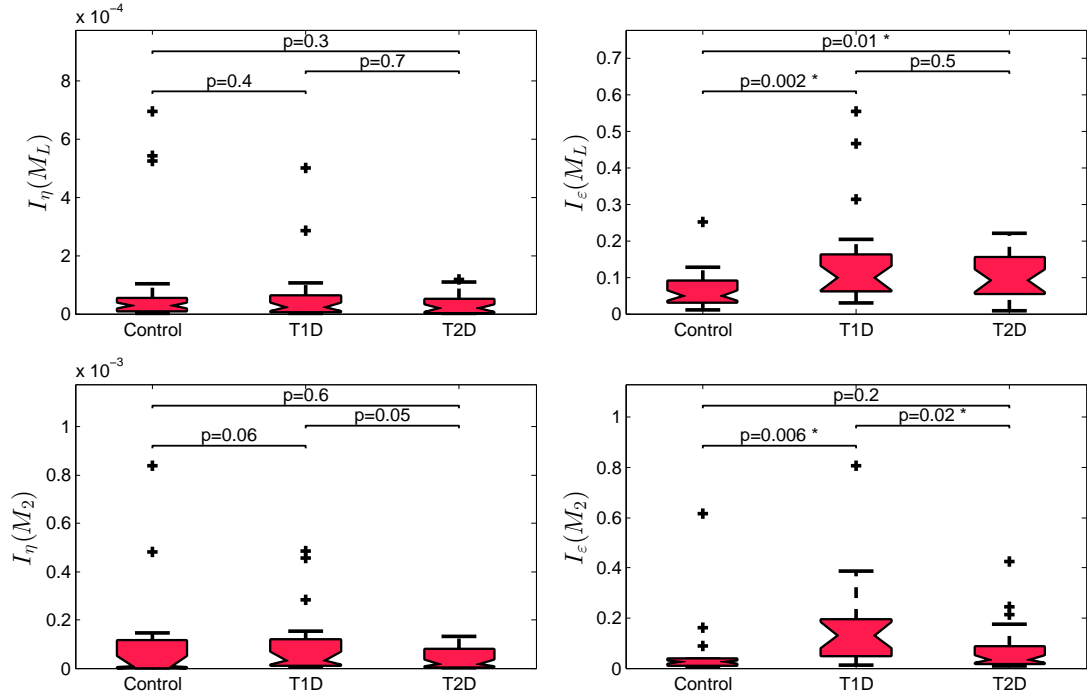


FIGURE 3.10: Boxplots for intensities of system noise I_ϵ in (a) M_L and (c) M_2 , and of measurement I_η noise in (b) M_L and (d) M_2

Note that when $\theta_{k_0}^2 < 4\theta_{k_1}$, the eigenvalues are complex numbers, and the fixed point is a spiral in the phase portrait.

3.4.3.3 Noise

The parameters of the stochastic terms, i.e the intensities of the system noise I_η and measurement noise I_ϵ in M_L and M_2 , are shown in Fig. 3.10. The intensities of the system noise are of the same order across all three groups for both models without any statistically significant difference: $p > 0.05$. The source of the system noise is the aggregated force accounting for other external factors such as physical activity, stress etc. which are not part of the model (only the impact of food intake on blood glucose dynamics is considered). The inclusion of system noise into the analysis provides more flexibility in the model's structure, but also accounts for its imperfections.

The measurement noise mainly comes from the inaccuracy of the readings from the CGM devices. There are significant differences in I_ϵ values between the control group and both DM groups for M_L . (Fig. 3.10 (b)). The higher intensities of measurement noise are observed in T1D and T2D groups than in controls ($p = 0.002$ and $p = 0.01$

correspondingly). This is well justified as the peak blood glucose levels are generally higher in the DM groups than in the control group, and the CGM devices are less accurate at the larger values. For M_2 , there is significant difference in measurement noise intensities between T1D and the other two groups.

3.4.4 Impulsive force

As discussed before, food intake is simplified as a bolus injection of glucose at $t = 0$. The value of F is an aggregated food impact, which is influenced by the quantity and the constituents of the food intake. The interpretation of F is difficult. The absorption of glucose is a slower process than an impulse function, and according to Dalla Man's maximal model [135], two compartments in the stomach and one in the intestine are involved in the absorption of glucose. It is difficult to directly link the value of F with either of the three compartments in the maximal model. Complex carbohydrates including polysaccharides (found in wheat, rice, potatoes, maize) require digestion into glucose prior to absorption into the blood stream, a process that takes significant time. Food that contains more liquid or solid would have a different glucose absorption rate. Glucose in liquid form is absorbed faster compared to the solid food, which may be reflected by a higher value of F as an initial impulsive impact. The slower and two different absorption rates for solid food, which could potentially be caused by a biphasic gastric emptying rate, may cause a double peak as described in Fig. 3.4(a) and a lower value of F . The delayed response in glucose uptake cannot be shown directly from the value of F since the exact meal time is unknown, but a combined consideration of the shape of the peak, the parameter values of the model together with the food impact can provide useful information about the relationship between the consumed food and the corresponding glucose response.

The comparison of the initial force parameter F among the groups is shown in Fig. 3.11. The median value of the food impact F is significantly larger in the T1D group compared to the control group. The larger and wider range of F show inadequate glucose regulation in the T1D group which may be caused by an inaccurate amount and time of the insulin injection by the patient or by the insulin pump. The median value of the food impact F for the T2D group is not significantly different from the control group, showing a better glucose regulation compared with the T1D group, which may suggest

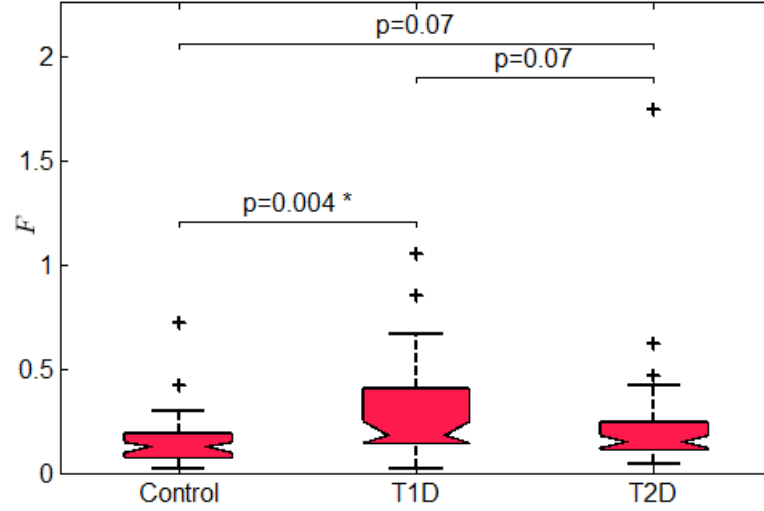


FIGURE 3.11: Boxplots for the initial force parameters F compared between the three groups.

that the medication taken by the T2D subjects was effective (refer to Table. 3.1 for details of medication used). According to [108], the time of taking hypoglycaemic drugs by a patient before the meal and the duration time of being on medication could also have an influence on F .

To investigate the effect brought by the hypoglycaemic drugs, a comparison has been made between five T2D subjects that take hypoglycaemic drugs regularly and a newly diagnosed T2D subject (No. 15) with no medication. The values of F obtained from all of the peaks of a single patient are grouped and represented by a box in a boxplot. All six subjects in the T2D group are shown in Fig. 3.12 (b). The values of F for subject No. 15 were compared with each other subject pairwise using the Wilcoxon rank sum test. The median value of F for the newly diagnosed patient who had not taken any medications ($F_{median} = 0.30$) was significantly higher ($p = 0.002$) than the median value of F across all the other five T2D patients ($F_{all\ median} = 0.12$) as shown in Fig. 3.12 (a). The lower F values in subjects who took regular medication imply that the suppression of the impulsive food impact could be a long-term effect of the hypoglycaemic drugs. Thus, the F value has the potential to be used as an indicator of how well the patient is managing the disease.

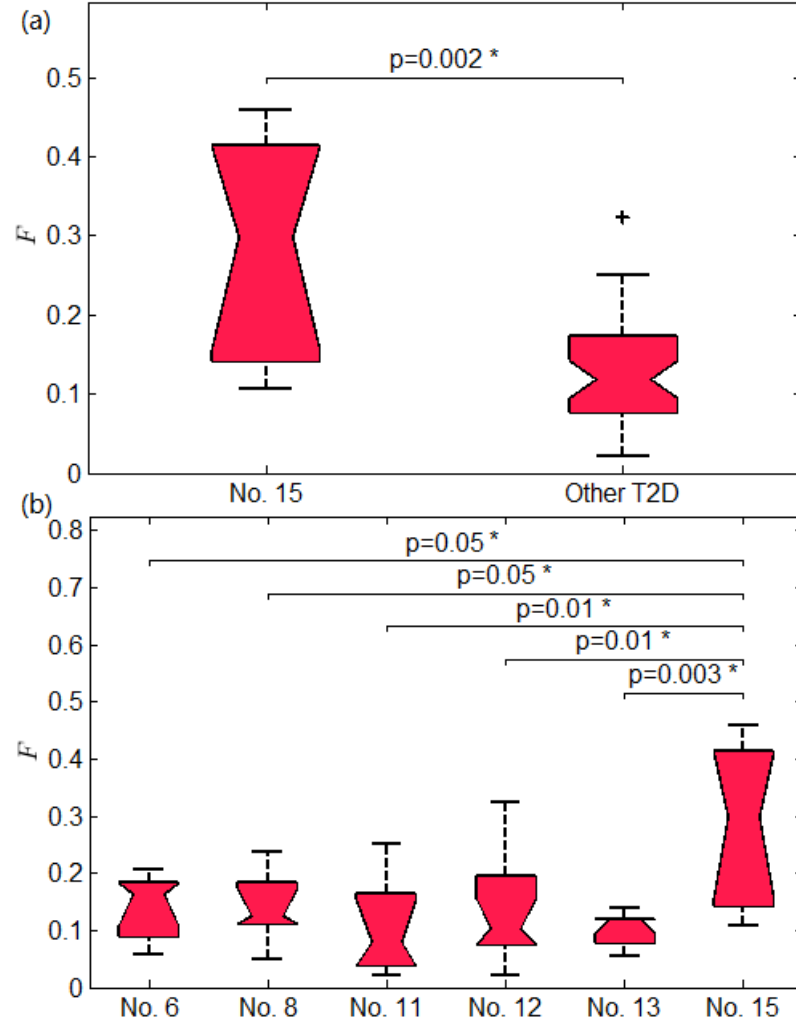


FIGURE 3.12: Boxplots for the food impact force F compared (a) between patient No. 15 and the rest of the T2D group; (b) among the subjects in the T2D groups.

3.4.5 Link between the data-driven model and physiological models.

The signs of pre-DM

As introduced in Section 3.1, physiologically based models are widely accepted in glucose dynamics modelling. Among them, the compartmental maximal model (MM) suggested by Dalla Man et al. [135] was designed to simulate the postprandial transient responses to food intake for subjects with and without DM. Twelve differential equations and thirty five parameters are included in the model MM . Different time series representing different subjects can be simulated by changing the values of the 35 model parameters. Three postprandial glucose responses characterising three non-DM subjects were simulated using the Simulating the Glucose-Insulin Response script provided in SimBiology

toolbox [155]. The standard parameter values suggested [135] were used: (i) without any signs of DM, (ii) with low insulin sensitivity, and (iii) with impaired β cell function. Cases (ii) and (iii) describe *potential* T2D patients with partially impaired pancreatic function, and are considered as pre-DM cases. The impaired function is compensated by either secreting more insulin or increasing tissue sensitivity, and thus keeping the blood glucose levels still within the healthy range. Note that M_L and M_2 belong to a different class of model compared to MM . As a physiologically based model, the glucose variation is only one of the variables that is generated by the MM using a set of population based parameter values that describe a typical pre-diabetes subject.

Treating these three simulated peaks the same way as the measurement time series in our cohort, the VB method was used to select between M_L and M_2 and infer the corresponding parameters. The purpose of inferring simulated time series is threefold. 1) If either M_L or M_2 can be fitted well to the simulated time series, the M_L and M_2 are proved to be consistent with the well-recognised compartmental MM . 2) The differences in the inferred parameter values between three simulated subjects can reveal how the parameter values change when a healthy person starts to develop T2D in theory. 3) A simulated time series is free from the limitations brought about by using the CGM devices, including a 3–12 minute delay from measurements and actual blood glucose level, measurement noise, and a fixed sampling frequency of one data point every five minutes.

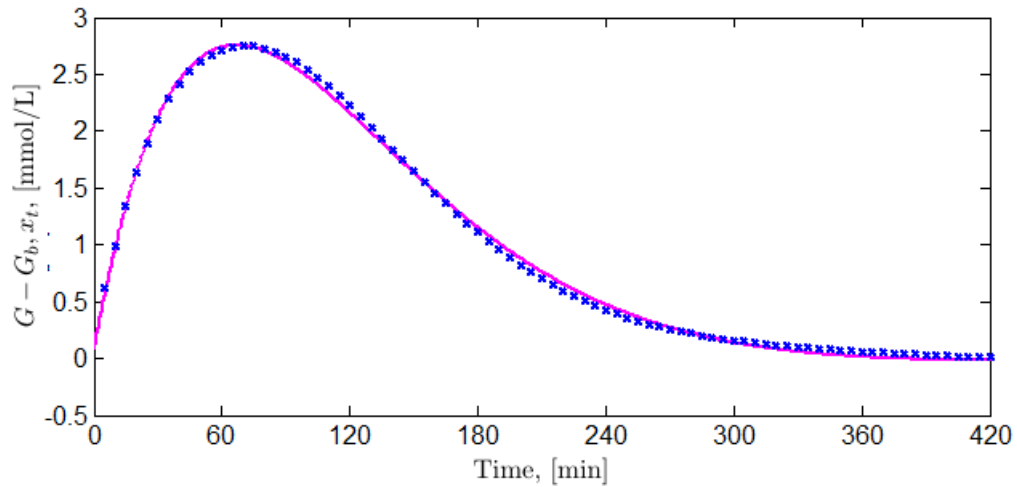


FIGURE 3.13: The dotted line is the simulated time series from the maximal model for a non-DM case without any signs of DM and the solid line is the deterministic solution using the parameter values inferred from model M_L .

All three simulated time series can be fitted by the linear model M_L , and an example result of fitting for case (i) is shown in Fig. 3.13. Values of the inferred parameter θ_k^{MM} are as follows respectively: (i) 0.027, (ii) 0.023, (iii) 0.021 [min^{-1}], and the values of θ_1^{MM} are (i) 0.4, (ii) 0.25, (iii) 0.16 [min^{-2}]. They have been compared with θ_k and θ_1 obtained for all measured peaks fitted by M_L for our cohort of subjects. The values of θ_k for all the peaks in the control group that modelled by M_L are presented as the left box in the boxplot shown in Fig. 3.14 (a), and the values of θ_k for all the peaks in the T2D group modelled by M_L is presented on the right box in Fig. 3.14 (b). The value of θ_k^{MM} for the simulated non-DM subject locates in the interquartile range of the box for the control group and in the top quartile of the box for the T2D group. The values of θ_k^{MM} for both pre-DM cases locate in the lower part of the box for the control group, but within the top part of the interquartile of the box for T2D. A qualitatively similar result is observed for parameter θ_1^{MM} (Fig. 3.14 (b)): the simulated value for case (i) is within the interquartile of the control group distribution and beyond the interquartile of T2D. Similarly, θ_1^{MM} for both pre-DM simulated subject locate within the bottom quartile of the control group and within the interquartile of the T2D range. This clearly shows that these two simulated pre-DM cases fall into the area between the non-DM and T2D distributions. It is worth noting that this observation is based on three simulated glucose variation time series, and it is too early to draw any significant conclusions without further validation from glucose dynamic data of pre-diabetes patients. This was an outcome that arose though the modelling performed. However, a close observation of the trend of the parameters θ_k and θ_1 might provide crucial information for early diagnosis of DM, particularly if such trends identify early abnormalities in glucose dynamics before the rise in the blood glucose concentration is considered significant.

This result provides same evidence of the robustness of the model (3.1a–3.1b) and its validation by comparison with an established phenomenological model [135]. The deterministic solution of our model (3.1a) in Fig. 3.13 indicated by the solid line contains only *two* parameters and matches the dynamics of the time series simulated using the *MM* [135] with *thirty-five* parameters indicated by the dotted line in Fig. 3.13. Being data-driven, our model takes full advantage of the CGM data, and, at the same time, reflects the intrinsic characteristics of the glucose-insulin system without detailed knowledge of the underlying physiological mechanisms.

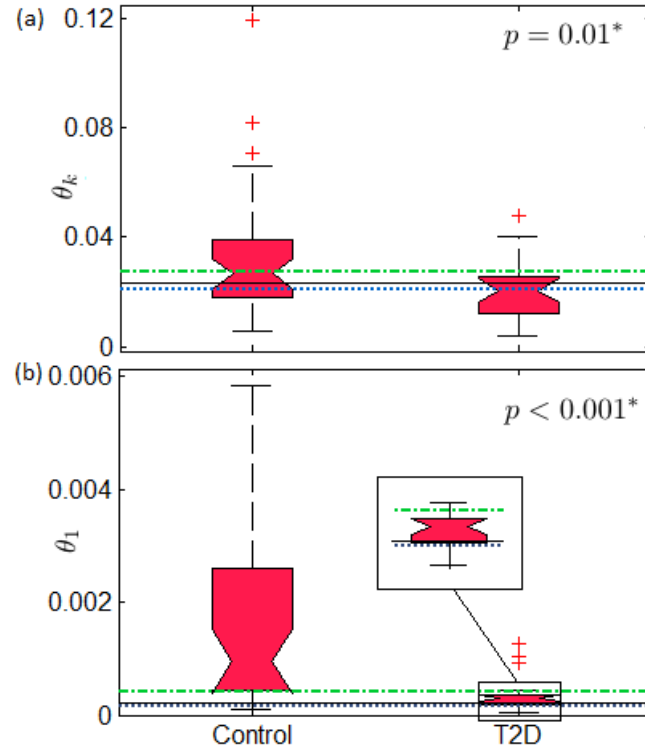


FIGURE 3.14: Boxplots for (a) θ_{k_0} and (b) θ_1 for all measured peaks fitted by M_L in our cohort of participants. Horizontal lines mark $\theta_{k_0}^{MM}$ and θ_1^{MM} for no signs of DM (upper dashed green line), low insulin sensitivity (middle solid line) and impaired β -cell function (lower dashed pink line) cases.

3.5 Conclusions and limitations of the study

The VB method introduced in Chapter 2 was successfully applied to develop dynamical models for transient glucose responses towards food intake and to infer model parameters for a cohort of fifteen subjects with or without DM. The results demonstrated a universal nonlinear stochastic model that is capable of capturing the dynamics of postprandial blood glucose excursions in a total of 132 peaks. The inferred deterministic parameters belong to different ranges for people with and without DM, demonstrating the potential for useful clinical applications including early diagnosis of DM. The parameter values were compared with three simulated subjects, two with pre-DM and one without DM, using the well-recognised MM , which allows useful physiological parameter interpretations. The significant difference in the food impact parameter values between a newly diagnosed T2D subject (without medication) and the rest of the subjects in the T2D group (with medication) implies that the food impact parameter may serve as an indicator of the impact of various drugs on the stability of blood glucose responses as well as control of DM.

This study is limited by a relatively small number of subjects, but the statistics presented are based on 132 peaks, and therefore is deemed sufficient to make conclusions. The results confirmed previous findings [108] suggesting an ability to distinguish DM and non-DM cases on the basis of model parameters, with promising interpretations for clinical use. The relationships that have been uncovered between model parameters and clinical group require further investigation to confirm these associations and elucidate the underlying mechanisms.

It is hoped that the developed modelling framework would be useful for other researchers to (i) successfully identify the parameters of their predictive models for blood glucose control, (ii) account for nonlinearity and stochasticity of the underlying process, (iii) consider the uncertainty in parameter estimations by using probabilistic parameter distributions rather than fixed values, and (iv) tackle the personalised needs of patients by considering individual, not averaged, time series.

In future, this study will be continued by including other factors that influence blood glucose variations, such as exercise, stress, etc. More subjects will be involved in the research within a controlled environment where the time and the macro nutrients of the food intake will be recorded in detail, which would potentially improve the model by decreasing the uncertainties in the delay between the food intake and the glucose response. A different input, such as a train of delta-functions, will be considered to incorporate the various gastric emptying rate into consideration. With all factors included in the model, a more precise description of the data is expected and a predictive model accounting for multiple external inputs could be developed.

Chapter 4

Post-transplant antibody dynamics

Antibody dynamics after kidney transplantation are of great clinical interest as they are considered to be associated with short and long term clinical outcomes [156]. However, the limited data on post-transplant antibody dynamics and their diverse behaviours have made the task of modelling difficult. There is no model available in the literature to analyse personalised dynamic patterns based on real patient data. In this chapter, a data-driven model has been developed for the first time to describe the evolution of antibodies in the critical first several months after transplantation. The VB method introduced in Chapter 2 has been applied to select the best model among the models with different orders and infer their parameters on a subject by subject basis. This work has been published by Zhang et al. in [157] and [158]. The structure of the chapter is as follows: Section 4.1 gives a background introduction to kidney transplantation and post-transplant antibody dynamics. Section 4.2 describes the data and presents visual analysis of the variety of dynamic responses of antibodies to transplantation. Section 4.3 explains the method used for model selection and parameter estimation. Section 4.4 presents the chosen model and detailed analysis of the model parameters. Section 4.5 summarises the results, outlines the relevance of the model for kidney transplant management, and justifies the need for further work.

4.1 Introduction

Kidney transplantation has been proven to be the best treatment for renal failure. Success of the transplantation is dependent on the reaction of the immune system primarily against *human leukocyte antigen* (HLA) of the transplant [159]. HLAs are proteins that help the immune system to distinguish between the body's own antigens and non-self antigens, and can be found on most cells in the body [160]. HLAs are encoded by a group of HLA genes allocated on chromosome six. A position on the chromosome is called a genetic locus. For humans, two alleles (an allele is a variant form of a gene) are located at each genetic locus, with one allele inherited from each parent. Based on the genetic loci, HLA genes are categorised into three groups: class I (HLA-A, -B, -C loci), class II (HLA-DP, -DM, -DQ, -DR loci), and class III (not the main concern for this study). The HLAs have more than 10,000 HLA Class I and over 3,600 HLA Class II alleles identified [161], therefore it is unusual to find two unrelated individuals with the same HLAs. Ideally, a recipient of the kidney transplantation must *fully match* the HLA proteins with the donor so that the immune system of the recipient does not trigger a harmful antibody response against the transplanted organ. In the early days of kidney transplantation, strict HLA matching rules required a perfect match at the -A, -B and -DR loci [162–164]. With the improvements in the graft rejection rate, the rules have been relaxed over the years [165, 166]. In the UK, only a minority of the transplants in the UK are fully matched for HLA [167] because of the shortage of fully matched organ donors.

One of the biggest problems of HLA-mismatched transplantation is the immune response of antibodies produced by B cells (a type of white blood cell) against the transplanted organ. Antibodies can come in different varieties known as isotypes. Out of these isotypes, Immunoglobulin G (IgG) is deemed to be the most detrimental to transplantation outcome [168]. When directed at a donor's HLA in the newly transplanted organ, IgG is referred to as *donor specific antibody* (DSA). DSAs are usually caused by a patient's previous exposure to non-self antigens via blood transfusion, pregnancy, organ or tissue transplant [169]. DSAs can form and persist for years after transplantation, and are frequently found at the time of transplant failure (when the graft loses the filtering ability to deal with the waste product from the blood) [170–172]. Post-transplant production of anti-HLA DSAs is indicative of an active immune response, and therefore increases

the risks of *Antibody-Mediated Rejection* (AMR) leading to a decline in renal function [173]. AMR is clinically confirmed by a renal biopsy (indicators are peri-tubular capillaritis, glomerulitis, peri-tubular capillary C4d staining) [174]. AMR includes immediate AMR (minutes after operation), acute AMR (within the first 30 days after operation, our focus in this research), and chronic AMR (developed over years). Some acute AMR can progress to a chronic phase resulting in eventual graft failure [175]. However, generally the long term consequences of the acute AMR, particularly episodes of lower severity, on graft function are uncertain [176]. In recent years, a number of publications [172, 177, 178] have confirmed that DSAs are the major cause of acute AMR and chronic graft failure. However, the association between acute AMR or chronic graft failure and high DSA levels can vary between patients. In the acute setting, transplantation across very high DSA levels may result in 50% graft loss, but data based on the currently used antibody detection assays cannot reliably predict the outcome [179]. Likewise, for chronic AMR, the relationship between the occurrence of AMR and the detection of circulating DSA is not clear [180, 181].

There are two type of DSAs: *de novo* DSAs and preformed DSAs. *De novo* DSAs do not pre-exist but develop after transplantation against the foreign graft, usually years after the transplantation [182]. They can cause a gradual damage to the organ that eventually results in organ failure [182–184]. Preformed DSAs, on the other hand, are present before the transplantation and thorough measurements of preformed DSAs are carried out to predict the likelihood of finding a compatible donor, and to avoid transplantation from a donor presenting HLA antigens to which the patient is sensitised [173]. Even with a low level of preformed DSA, graft outcome may be compromised [185]. Transplantation that is performed across an HLA barrier with preformed DSAs present is defined as *antibody incompatible transplantation* (AiT) [175]. To allow AiT, there is a consensus on removing DSAs before transplantation for patients with DSA levels above a predetermined threshold, which has not been determined systematically [174]. Clinically, it is difficult to eliminate preformed DSAs completely because the removal of anti-HLA DSAs has generally been followed by re-synthesis of DSAs. Repeated treatments may create a period of time during which the anti-HLA DSA levels drop to an acceptable level, allowing for the transplantation to be undertaken. To remove preformed DSAs immediately before transplantation, different methods, such as plasmapheresis, double filtration, immunoabsorption, have been developed [156, 159, 174, 186, 187]. Since the

incorporation of plasmapheresis, the acute AMR occurrence rate has dropped to about 5 – 7% of all kidney transplantation [188]. After the transplantation, because of immunological memory, an immune response can be triggered towards the graft by re-synthesis of the anti-HLA DSAs that have been removed prior to the transplantation, resulting in severe acute AMR and an increased risk of graft loss [174]. Therefore, safe transplantation of potential recipients with high levels of circulating DSAs, i.e. highly sensitised recipients, is an ongoing problem resulting in prolonged waiting times for transplantation [159]. Such a type of transplantation is defined as a high risk kidney transplantation in clinical practice.

It has recently been recognised by the transplant community [156, 189] that post-transplant screening for anti-HLA DSA could be an important tool for monitoring of transplant recipients. Early DSA dynamics are likely to profoundly affect clinical outcomes [156, 190, 191]. It is observed that the dynamic behaviour of post-transplant DSAs varies from case to case, and even different DSAs in the same patient (targeting different HLA) show diverse patterns. A strong mathematical approach to describe the dynamics of the preformed DSA is in need. It can help clinicians, e.g. by suggesting what laboratory assays can be developed to detect more detrimental DSAs, and the time points at which laboratory assays should ideally be collected. Once appropriate assays are available, the modelling may help in the interpretation of results of the assays at different time points. This could be particularly important in relation to falls in DSA levels, since this is a key clinical objective that is currently not achievable in clinical practice. However, no such mathematical model exists.

Physiologically based models are not yet feasible to analyse the antibody dynamics due to the complexity of the underlying immunological responses to transplants. The possible mechanisms underlying changes in DSA levels are complex. DSA levels may change because of rises and falls in the rate of production which are controlled by multiple factors. Falls in the levels of DSA post-transplant are very interesting, as they may occur much faster than the ‘natural’ rate of antibody clearance from the body — antibodies are estimated to have a half life of about 20 – 30 days [192]. Mechanisms associated with reductions in antibody levels could include the absorption of antibodies onto HLA molecules on the graft [182]— it is known that the levels of HLA on a graft may increase post-transplant, but this cannot yet be quantified. Some HLA is shed by the graft so antibodies could be absorbed in the circulation. It is known that one physiological

method used by the body to control antibody levels is to produce antibodies that block other antibodies (idiotypic antibodies), and the production of idiotypic antibodies could explain the falls in DSA post-transplant [193]. However, as with other potential regulatory mechanisms, it is currently hard to measure idiotypic antibodies accurately, and therefore, such a regulatory mechanism hypothesis based on the production of idiotypic antibodies cannot be proved experimentally under current techniques.

Data-driven models, on the other hand, require an accurate way of measuring the DSA in human sera, which was not possible until very recently [194]. Due to the development of Luminex technology based on single-antigen bead assays in recent years [156, 190, 195], these highly sensitive and specific assays using HLA proteins meet the increasing need for monitoring post-transplant DSA [196] and open up opportunities to develop data-driven mathematical models for the evolution of antibodies after transplantation.

A unique dataset with detailed antibody measurements using the single-antigen beads assays spanning three to six months has been obtained by our group [156]. A previous analysis [156] of this data set revealed various patterns of antibody dynamics, both with or without acute AMR. Some DSA time series show a rapid rise during the first two weeks followed by a rapid fall to almost undetectable levels, which then remain low. This finding is striking: in many of these patients, the DSAs had persisted for many years before transplantation, and therapies used experimentally have been unable to stop antibody production before transplantation. A better understanding of this phenomenon could therefore have practical benefits.

The aim of this work is therefore to describe early antibody response within the first six months of transplantation in mathematical terms by deriving appropriate dynamic models and analysing the dynamic responses in relation to the occurrence of acute AMR. As a key early outcome in AiT, acute AMR episodes are associated with the levels of immunosuppression required in the post-transplant period, and are also associated with short and long term graft survival [176]. This approach might enable a more intelligent application of laboratory testing and suggest therapeutic approaches to selectively control this antibody response and improve clinical transplant outcomes.

4.2 Data description and visual analysis of dynamic patterns

Data from twenty-three patients who underwent renal AiT at University Hospitals Coventry and Warwickshire (UK) (UHCW) between 2003 and 2012 were analysed in this study. The data comprised of time series of DSA evolutions over a period of about ten days before and six months after transplantation. Serum samples for DSA analysis were taken almost daily in the first three to four weeks, as most dynamic behaviour occurs during that period, and sampling became more sparse later when the antibodies tended to be more stable. Antibody levels were measured using the microbead assay manufactured by One Lambda Inc (Canoga Park, CA, USA), analysed on the Luminex platform (XMap 200, Austin, TX, USA). The assay measures the Mean Fluorescence Intensity (MFI) which corresponds to antibody level although the relationship is linear only over a limited range. It is known that the inter-assay coefficient of variability for DSA measurements is around 10-30% [197]. As described in [156], when the MFI value is higher than 10,000 AU (Arbitrary Units) and below about 1,000 AU, the linear correlation breaks.

Some of the patients had multiple DSAs targeting different HLA, so the total number of post-transplant time series available for this analysis was *thirty-nine*. *Twenty-seven DSA time series* belonging to fourteen patients that experienced episodes of acute AMR in the first thirty days after transplantation (AMR group), and *twelve DSA time series* belonging to the other nine patients who did not have an episode of AMR (no-AMR group). Rejection episodes were diagnosed by renal biopsy or clinically if there was rapid onset of oliguria with a rise in both serum creatinine and DSA levels [156]. In patients receiving HLA antibody-incompatible grafts, the incidence of AMR was 30 – 40% [191]. Although AMR can be severe and can eventually result in graft failure, it usually develops slowly over a period of several days. This gives an opportunity to detect AMR at an early stage and treat it, resulting in better outcomes [156, 190].

Visual examination of the time series reveals diverse dynamic behaviour of DSAs. Fig. 4.1 and Fig. 4.2 show some examples of the patterns from the no-AMR group and the AMR group respectively. As some patients had multiple DSAs, the case number in these figures and in the corresponding text is followed by the DSA type. For example, in Fig.

4.2, patient 36 had two anti-HLA DSAs, HLA-A24 and HLA-DR17, comprising two different time series: case 36 HLA-A24 (case 36 A24 for short) and case 36 HLA-DR17 (case 36 DR17 for short). Pretransplant antibody removal (before day 0) can be seen to reduce total DSA levels due to cycles of double filtration plasmapheresis. Typically, between two and five alternate day sessions were performed.

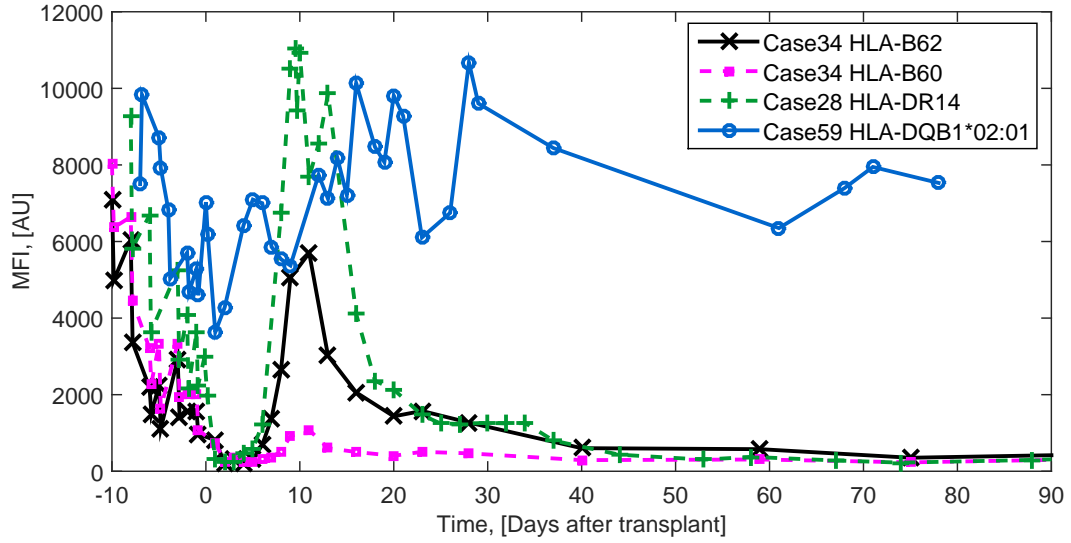


FIGURE 4.1: Measured time series illustrating individual DSA changes in the no-AMR group. Markers correspond to each measurement point. MFI=mean fluorescent intensity.

The initial drop is typically followed by a rapid rise in DSA which usually occurs with a lag of a few days after transplantation (day 0) and is caused by two factors: plasmapheresis stopping and an increased rate of DSA synthesis due to an immunological memory response. After the peak levels a diversity of dynamic patterns is noticeable: antibody levels do not follow a common route, varying from case to case, and even differing for different DSAs in the same patient. In some cases there is a rapid fall in DSA to a steady state, corresponding to a low (almost zero) level of DSA, and this is typically reached within the first month after operation. Such patterns are observed in both the no-AMR and AMR groups: case 34 (both B62 and B60) and case 28 in Fig. 4.1, and case 36 DR 17 in Fig. 4.2. In other cases the dynamics of the fall after the peak are followed by another rise, and antibodies do not settle at a low level within the first three months after operation: case 59 in Fig. 4.1, case 36 A24 and cases 61 and 69 in Fig. 4.2. They either demonstrate a slow dynamic around a certain constant level (case 61 in Fig. 4.2) or change dramatically over the first three months (case 59 in Fig. 4.1 and case 69 in Fig. 4.2). There is no obvious relationship between these dynamical patterns,

steady state levels and the occurrence of AMR episodes. In some cases, as shown above, low steady state levels are observed in the no-AMR group and higher levels or dramatic changes are noticeable in the AMR group. There are also cases of the absence of AMR despite high levels of DSA or presence of AMR despite low DSA levels. Finally, some patients (e.g. case 36 in Fig. 4.2) rejected the kidney, but had multiple DSAs with one type that rose after the initial fall post-transplant (A24) and another type that kept falling to a low steady level (DR17). This visual analysis demonstrates that there is no certain association between higher levels of post-transplant DSA and the occurrence of the rejection episodes.

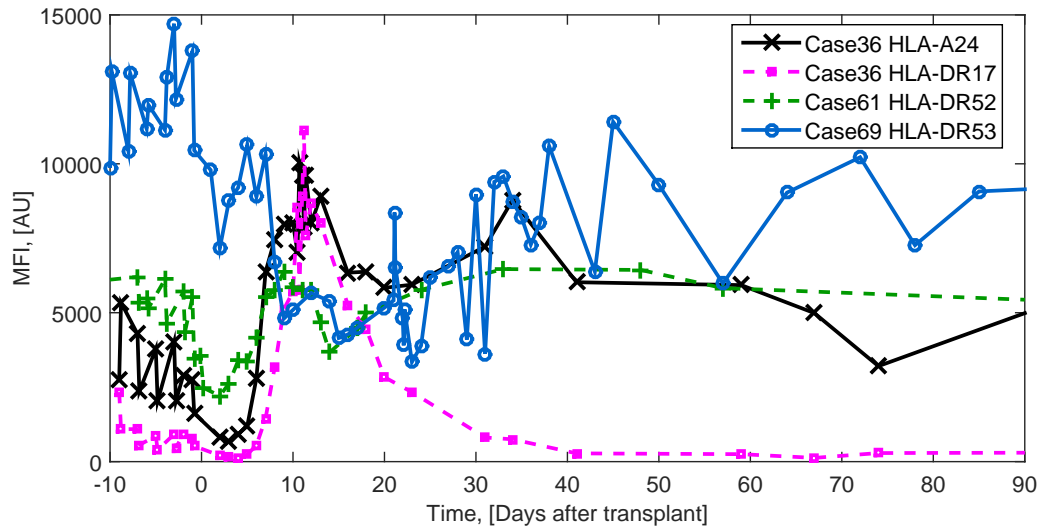


FIGURE 4.2: Measured time series illustrating individual DSA changes in the AMR group. Markers correspond to each measurement point. MFI=mean fluorescence intensity.

The aim of this study was to analyse these dynamical patterns in order to propose a set of characteristics capable of discriminating between the patients with and without the incidence of AMR. In this study, we are particularly interested in the DSA dynamics after the first peak value down to an almost zero level, i.e. focus is on the typical pattern of a rapid fall that occurs in most of the patients with and without AMR episodes. Falls in the serum levels of anti-HLA DSA after kidney transplantation are of great clinical interest, as they are associated with resolution of rejection and good long term outcomes in patients at high risk of graft loss [191].

4.3 Models and methods

4.3.1 Data fitting and model selection

As seen from the preliminary observations of the dynamic patterns, the anti-HLA DSA response to the transplanted kidney is a complex immunological process, nonlinear and stochastic in general. Time series available for analysis are complex and one-dimensional: only one variable as a function of time (MFI levels) is available representing a response of the entire immune system to the external stimuli (a newly transplanted kidney). A mathematical model can quantitatively describe the dynamics in the post-transplant DSA time series and capture the main features that are shared between different DSA time series. These diverse dynamic patterns pose a set of challenging questions with respect to the form of the function, the order of the system and the number of parameters to be used in the model. It is also unclear whether the system equation should be linear or nonlinear, stochastic or deterministic, and what would be the most appropriate modelling approach to identify system parameters in the situation where no preliminary knowledge of the model is available. Although we only consider the falling part of MFI level dynamics, all the above questions remain.

4.3.1.1 Exponential fitting

It can be noticed that the falling MFI dynamics of HLA DSA after the peak value is a relaxation process, the simplest theoretical description of which is an exponential law. Initially the curve fitting tool (Cftool) in Matlab [198] was used to fit each of the thirty-nine DSAs. Some of the time series were described by this approach sufficiently well; however, the use of superposition of exponential functions could not describe all the cases with and without AMR in our cohort. As the next step, instead of exponential functions, i.e. *solutions* of linear dynamic equations, dynamic mathematical models in the form of differential equations were considered.

It is worth mentioning that there are other parameter estimation and system identification methods, such as the least squares method, available, especially for linear systems. However, the main purpose is to search for the best possible model that describes the data regardless of the form of the model – linear or nonlinear, deterministic or stochastic.

Therefore, the employed method needs to be able to compare between linear/nonlinear and deterministic/stochastic models. Classical system identification methods such as the least squares method would constrain the model choice within linear and deterministic forms. In addition, most commercially available toolbox such as the system identification toolboxes [199] in MATLAB cannot deal with data with highly irregular sampling frequency, which is indeed the characteristic of the data as described in Section 4.2. The DSA measurements 30 days after transplantation are highly irregular and sparse, as seen in Fig. 4.1 and Fig. 4.2. Therefore, Variational Bayesian method fit the purpose of the study perfectly and thus, it has been chosen as the prime method in this chapter, similarly to the previous application discussed in Chapter 3.

4.3.1.2 Form of the model: linear/nonlinear and stochastic terms

The general form of an n -th order nonlinear Ordinary Differential Equation (ODE) with coefficients in the form of a polynomial function, as proposed in Section 2.1, have been considered. Initially two stochastic terms were included to represent noise in the system equations. Measurement noise is added due to uncertainty in the measured data, and the dynamic noise accounts for any other hidden properties not captured by the model. Thus, DSA falls after the initial rise (to a peak level) in the early post-transplant period can be described by the following model:

$$\frac{d^n}{dt^n}x_t + \sum_{i=0}^{n-1} f_{i+1,\theta_i}(x_t) \frac{d^i}{dt^i}x_t + f_0(x_t, \theta_0) = \eta_t \quad (4.1a)$$

$$y_t = x_t + \varepsilon_t \quad (4.1b)$$

Equation (4.1a) is an evolution equation of n th order, where x_t is a function of t that describes the MFI dynamics, and y_t is the measured MFI time series. η_t is system noise, and ε_t is measurement noise. Each noise was modelled as Gaussian-distributed white noise with zero mean and intensity (variance) of I_η and I_ε respectively. $f_{i+1}(x_t)$ ($i = 0, 1, \dots, n-1$) are polynomial functions of x_t . The derivative of order zero of x_t is defined to be x_t itself. The order of the system equation n is to be decided together with the unknown parameters θ_i ($i = 0, 1, 2, \dots, n-1$). n initial conditions are required to obtain a closed form solution.

Model M_n constituting (4.1a) – (4.1b) covers a variety of dynamic patterns depending on the order of the system n . A more complex model may be able to explain a wider range of system behaviour in the data at the risk of overfitting.

4.3.2 Model and parameter identification

As explained in Chapter 2, both the model and parameter identification were carried out using the VB method. The freely available SPM9 toolbox [77] (referred to as the VB toolbox) for MATLAB [198] allows accounting for both types of stochastic terms: measurement noise and system noise.

Starting from the first order linear model M_1 ($n = 1$ in (4.1a)) where the coefficients $f_{i+1}(x_t)$ are constants with constant parameters θ_i ($i = 0, 1, \dots, n - 1$), the order n was increased until the model M_n fitted the data sufficiently well, i.e. satisfied the criteria given in Section 4.3.3. Attention has to be paid to the features in the dataset that can be explained by a model with a higher order, but cannot be explained by the model with a lower order, and to decide if the features are general enough to make the final decision on the order for all DSA time series under investigation.

For each model candidate M , free energy \mathcal{F}_M was maximised by iteratively optimising the states of the system and model parameters to approach the model log-evidence $\log(p(y_t|M))$, as introduced in Section 2.2.2. This procedure is embedded into the VB algorithm as introduced in Chapter 2 Section 2.3. The maximised value of the free energy, among other criteria (normalised root mean square error and the stability of the immune response, both of which are discussed in the next section), defined how well the model performs.

As explained in Section 2.5, the priors regarding the parameters are important. Such information on possible parameter values was not available to us, and therefore the mean values of the parameter priors were set to zero. To allow the algorithm to search in a relatively wide region for the optimal parameters, all variances were set to be 10^4 , i.e. priors with wide distributions were considered. Both noise precisions α_η and α_ε were modelled by a gamma distribution with two hyperparameters (shape a_η , a_ε and rate b_η , b_ε respectively). Weakly informative Jeffreys priors, as described in Section 2.5, were chosen for the precisions of the noise, with both shape and rate parameters set to 1. The

initial conditions were all modelled as Gaussian distributions. The prior means of the initial conditions were defined from the measurement time series, and the prior variances were set to 10^4 .

4.3.3 Model selection criteria

The following four criteria were applied to identify the best model.

1. *The free energy \mathcal{F}* has been maximised by tuning the system parameters in an iterative manner for each model. Note that decision making based on the comparison of the free energy of any two models with different orders could be problematic due to the heavy penalisation of the model complexity embedded in the VB method (as explained in Section 2.4). Increasing the order of the system by one would not only increase the degrees of freedom in the parameter space, but also increase the dimension of the system states. This leads to a dramatic decrease in the free energy, which could be an order (or several orders) greater than the free energy difference between models of the same order. Therefore, the free energy criterion was only used to compare models of the same order. For models with different orders, criterion 2, as below, was utilised.

2. *Normalised Root Mean Squared Error (NRMSE)* was used to compare the models with different orders for each individual time series. Inferred parameters θ were applied back to the system equation to generate time series without stochastic terms, i.e. the deterministic solution. Note that because parameters were identified in the form of normal distributions, the most probable (mean) values of the parameters were substituted into the system equation. The *root Mean Squared Error (RMSE)* between the measurement MFI time series y_t and the inferred deterministic time series \hat{y}_t can be calculated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}} \quad (4.2)$$

NRMSE accounts for the different heights of the peaks for each DSA time series and is found by dividing the RMSE by the maximal MFI value for a given DSA time series. The model with the lowest value of NRMSE describes the data most accurately. For the model to be deemed satisfactory, NRMSE should not exceed the value of 0.15 (or 15 %) as it is known that the inter-assay coefficient of variability for DSA measurements is

around 10-30% [197]. It could be argued that NRMSE only represents the goodness-of-fit without considering the model complexity, whereas the AIC (as introduced in Section 2.4) accounts for the complexity of the model by subtracting the total number of the model parameters from the logarithm of the mean squared error. In this analysis peak heights vary between 1000 AU and 10,000 AU, so the RMSE value of each time series needs to be normalised by its peak height before the normalized RMSE (NRMSE) values are compared across all 39 time series. Therefore, the NRMSE was chosen over the AIC in order to take different peak heights into consideration.

3. *Generic form.* As the entire aim was to find a model capable of capturing the common patterns in all time series, a model that could only describe some of the DSA time series was disregarded.

4. *System stability.* Due to the chosen dynamic pattern in the data where the antibody level falls rapidly to an almost zero level and remains low, the model describing such data has to have a unique stable steady state, which implies that the system's response decays with time. This has been checked via calculations of the real parts of the corresponding eigenvalues which have to be negative for stability. Note, even though the steady state of the immune homeostasis was disturbed by transplantation, the antibody levels settled rapidly to a new steady state except for the extreme cases (example case 069 HLA-DR53 in Fig. 4.2), but consideration of such cases is outside the scope of this work.

4.4 Results and Discussions

4.4.1 Model selection

To choose the best model that is capable of describing all of the falling dynamics of the DSA time series in the cohort, a series of choices needed to be made: 1) the order of the system, 2) deterministic (measurement noise only) or stochastic model (both system and measurement noise), 3) linear or nonlinear models. The results are presented in the following sections.

4.4.1.1 Comparison of linear models of different orders

Linear models with different system orders were considered first. The equation (4.1a) in Section 4.3.1.2 transforms into a linear differential equation when the coefficients $f_{i+1}(x_t)$ are constants, with constant parameters θ_i ($i = 0, 1, \dots, n-1$), where $f_n(x_t) = \theta_n, \dots, f_2(x_t) = \theta_2, f_1(x_t) = \theta_1$. $f_0(x_t)$ is defined as $-\theta_0$ for convenience.

A first order linear model was considered first, and it did not show good performance. Then linear models with higher system orders were investigated. In this section, we present the methodology of system and parameter identification by comparing solutions for linear first, second and third order dynamic equations only:

$$\text{Model 1 } (M_1): \quad \frac{dx_t}{dt} + \theta_1 x_t - \theta_0 = \eta_t \quad (4.3)$$

$$\text{Model 2 } (M_2): \quad \frac{d^2 x_t}{dt^2} + \theta_2 \frac{dx_t}{dt} + \theta_1 x_t - \theta_0 = \eta_t \quad (4.4)$$

$$\text{Model 3 } (M_3): \quad \frac{d^3 x_t}{dt^3} + \theta_3 \frac{d^2 x_t}{dt^2} + \theta_2 \frac{dx_t}{dt} + \theta_1 x_t - \theta_0 = \eta_t \quad (4.5)$$

Note if the third order equation had not been successful, the procedure would have continued to account for nonlinearities (presented in Section 4.4.1.3) first and then increased the order of the system until a suitable solution was found.

Initially not only the measurement noise ε_t (as in (4.1b)) but also the system noise η_t (as in (4.1a)) was included in the models. It was found that for all DSA time series, models without system noise have larger free energy compared with the counterpart mathematical representations containing both types of stochasticity (an example is shown in Section 4.4.1.2). The benefit - improved fitting - obtained by using the more complex model with system noise does not exceed the penalty introduced by adding two degrees of freedom in the parameter space. Therefore, we excluded the system noise from the models and this is reflected in the zero right hand side of (4.3) – (4.5).

Typical fittings for four DSA time series, one from the no-AMR group and the other three from the AMR group, by the three suggested models (4.3) – (4.5) are shown in Fig. 4.3. The results for models $M_1 - M_3$ in Fig. 4.3 (a) and (c) show a winning model candidate M_3 . Even though (a) is from a patient in the no-AMR group and (c) is from a patient in the AMR group, both time series show oscillations after day 30. M_1 failed to describe the dynamics of both time series as indicated by large NRMSE values in

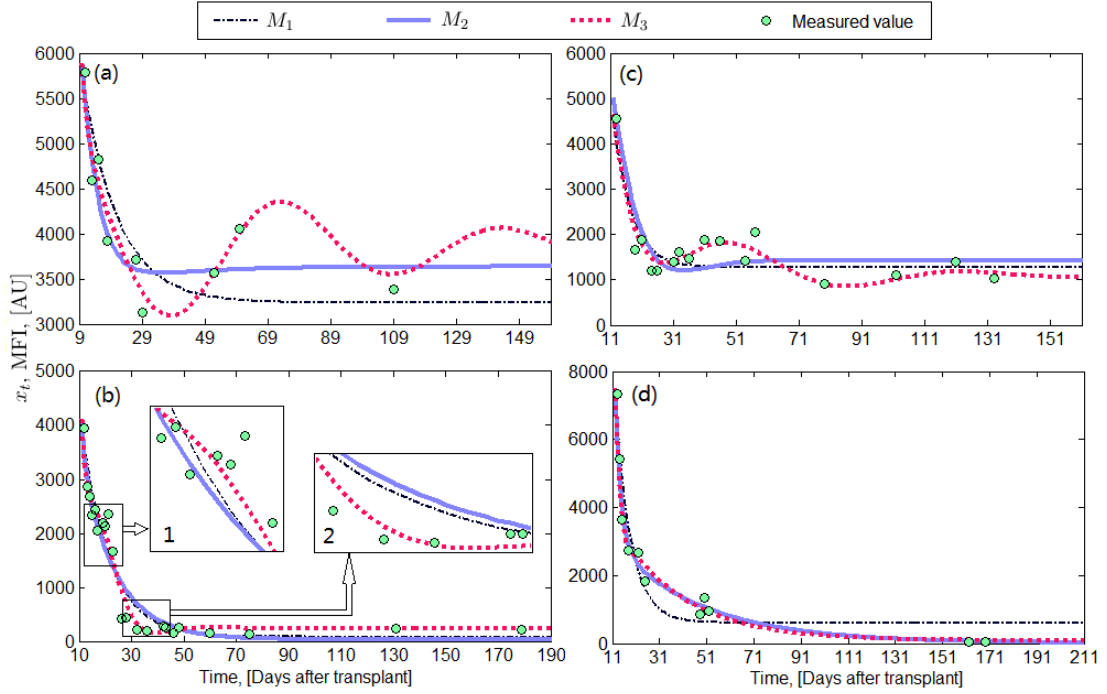


FIGURE 4.3: Typical fitting results compared among the three models $M_1 - M_3$ for (a) HLA-B60 (case 52) for a patient from the no-AMR group; (b) HLA-DRB3*01 for a patient (case 14) from the AMR group; (c) HLA-A32 for a patient (case 16) from the AMR group; (d) HLA-A2 for a patient (case 17) from the AMR group. The measured values are indicated by circles.

Table 4.1: NRMSE= 0.272 and NRMSE= 0.090. M_2 successfully described the initial falls for both time series, but failed to capture the oscillations in DSA after day 30, which is also confirmed by the large NRMSE value of 0.053 and 0.096 (Table 4.1). M_3 captured successfully both the falling part and the later trend with smaller NRMSE values of 0.014 and 0.053. Fig. 4.3 (b) exhibits different dynamics with a cluster of data around day 20. This is a common feature observed in the majority of time series in both AMR and no-AMR groups, and requires special attention. The temporary stall of falling could not be expressed by using M_1 or M_2 ; however, M_3 successfully depicted the sudden changes in falling as shown in the magnified box 1 in Fig. 4.3 (b). Further, the fittings by M_1 and M_2 were almost indistinguishable after day 70, and both models underestimated the settling level of the DSAs. The fitting by M_3 otherwise correctly estimated the settled value and gave a better description of another clustered region around day 30 (see the magnified box 2 in Fig. 4.3 (b)). Thus, M_1 and M_2 were ruled out based on their incapability of describing the important features, and the higher order model M_3 was chosen.

From Table 4.1, it is noticeable that the free energy values of a model with a lower order

TABLE 4.1: Summary of the free energy and the NRMSE values for three models of different order corresponding to the four example datasets in Fig. 4.3.

	Free Energy				NRMSE		
	M_1	M_2	M_3	M_3S	M_1	M_2	M_3
(a)	-99	-164	-233	-1036	0.272	0.053	0.014
(b)	-182	-346	-507	-634	0.083	0.085	0.013
(c)	-150	-268	-387	-1692	0.090	0.096	0.053
(d)	-110	-204	-292	-463	0.088	0.071	0.073

are consistently larger than the free energy values of a model with a higher order. A higher order model has a larger number of hidden states. For example, M_2 has twice as many hidden states as M_1 . Since the calculation of the free energy takes the inference of hidden states into account, the uncertainty of the larger number of the hidden states in the model with a higher order increases the K-L divergence, causing a lower free energy value. Therefore, NRMSE is a better criterion, compared to the free energy value, for model selection between models with different order.

The same approach was applied to all the other DSA time series. In 32 out of 39 cases, the NRMSE value of M_3 was the smallest among the three models. In the other 7 cases, the NRMSE value of M_2 was comparable with the NRMSE value of M_3 . An example is shown in Fig. 4.3 (d) where the fittings by M_2 and M_3 were indistinguishable from each other, with close NRMSE values as in Table 4.1. To compare the NRMSE values between M_2 and M_3 across all 39 cases, the NRMSE value of M_3 was subtracted from the NRMSE of M_2 , and the differences for all time series are shown in the boxplot Fig. 4.4. The differences in the NRMSE between the two models were tested by the one-sample t-test at the significance level of 0.001, and the mean value was found to be significantly larger than zero. Therefore, the deterministic model M_3 was selected as the best model across the cohort with the dynamic equation in the form:

$$\frac{d^3x_t}{dt^3} + \theta_3\frac{d^2x_t}{dt^2} + \theta_2\frac{dx_t}{dt} + \theta_1x_t - \theta_0 = 0 \quad (4.6)$$

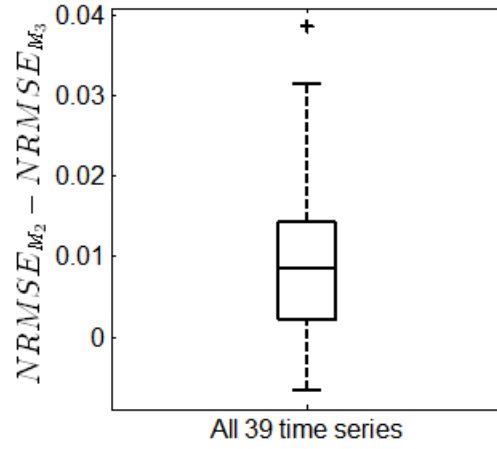


FIGURE 4.4: Boxplot of the difference between the NRMSE of M_2 and NRMSE of M_3 .

4.4.1.2 Deterministic versus stochastic

As stated in Section 4.4.1.1, there are two submodels for the model with the order of three: M_3 with measurement noise only, the other one with system noise and measurement noise, referred to as M_3S . A comparison example of the time series in Fig. 4.3 (b) between the fittings obtained from M_3 and M_3S is shown in Fig. 4.5. It is visually clear that M_3 fit the data better than M_3S , which can be confirmed by a higher free energy value of M_3 ($\mathcal{F}_{M_3} = -507$ as shown in Table 4.1) compared with the free energy value of M_3S ($\mathcal{F}_{M_3S} = -634$). Intuitively, M_3S is structurally more flexible to reflect the variations in the data by allowing large system noise intensity, i.e. with large enough system noise, M_3S can fit any data. Therefore the accuracy of the model improves, reflected by the smaller measurement noise intensity of M_3S ($I_\varepsilon(M_3S) = 132$ AU) compared with the noise intensity of M_3 ($I_\varepsilon(M_3) = 259$ AU). However, the deterministic fitting of M_3S is worse than M_3D , because M_3S loses important information carried by the data about the system by categorising it as system noise. Without enough measurement data, such miscategorisation is difficult to avoid. With two more degrees of freedom introduced in describing the system noise (two hyperparameters of the system noise), the complexity term of the free energy increases (as explained in Section 2.3.4), causing a smaller value of free energy for M_3S , in spite of a larger accuracy term. These results hold true for all of the time series. The free energy comparison between M_3S and M_3 for the other three examples shown in Fig. 4.3 can be found in Table 4.1.

Therefore, M_3 is chosen over M_3S for all of the time series in the cohort. However, when the measurements are sparse and corrupted with noise, the data have limited

constraining power over the parameters to support a more complicated model with system noise. The falling dynamic of DSAs, as a result of the complex immune response towards the newly transplanted organ, contains system noise. When more and better data with less measurement error become available, the information about the system noise may be extracted from the data.

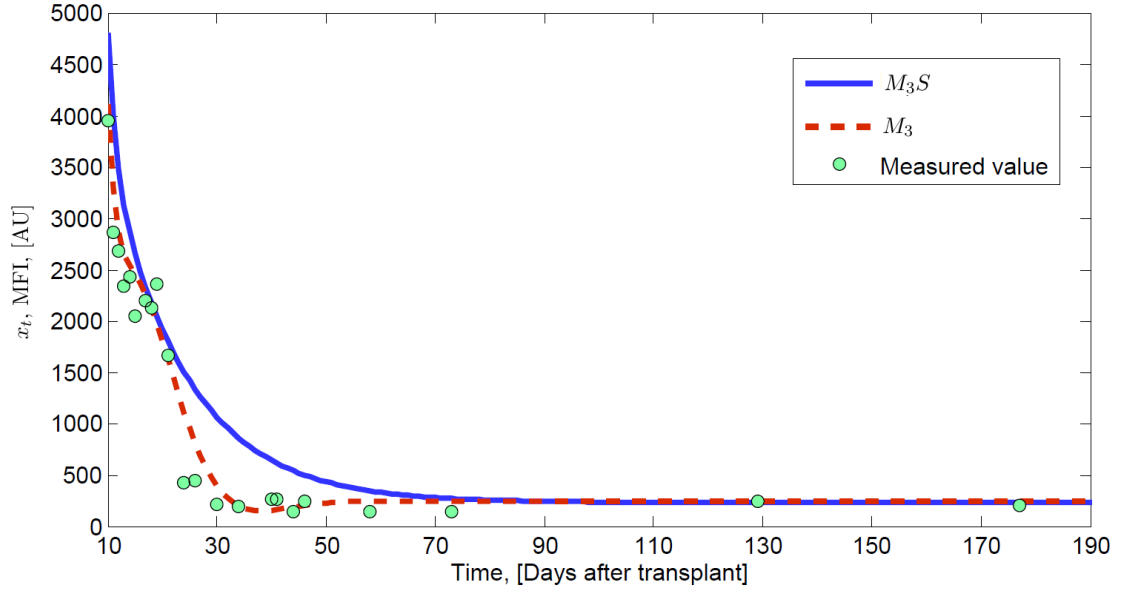


FIGURE 4.5: Fitting comparison between M_3S and M_3 for the time series in Fig. 4.3 (b).

4.4.1.3 Nonlinear versus linear

Polynomial forms allow a description of a wide range of nonlinear solutions, and can be adjusted to fit different dynamic features by increasing the number of components and by varying their parameters. To keep a parsimonious form of the system equation, nonlinearity was introduced into the second order model (the lowest possible order considering the first order does not provide good fitting), in the form of polynomial nonlinear coefficients $f_1(x_t)$ or $f_2(x_t)$ in (4.1a). It was acknowledged that a linear description is preferable over nonlinear if this does not increase the number of unknown parameters dramatically. Consequently, the maximal number of unknown parameters in the second order nonlinear model was kept comparable with the number of parameters of the third order linear equation, i.e. no more than 4. Under this condition, two nonlinear models were considered: model NM_1 with nonlinearity in the damping term ($f_1(x_t) = k_1x_t + \theta_1$, $f_2(x_t) = \theta_2$) and model NM_2 with nonlinearity in the x_t term

($f_1(x_t) = \theta_1$, $f_2(x_t) = k_2x_t + \theta_2$). The corresponding system equations are as follows:

$$NM_1: \frac{d^2x_t}{dt^2} + \theta_2 \frac{dx_t}{dt} + (k_1x_t + \theta_1)x_t - \theta_0 = 0 \quad (4.7)$$

$$NM_2: \frac{d^2x_t}{dt^2} + (k_2x_t + \theta_2) \frac{dx_t}{dt} + \theta_1x_t - \theta_0 = 0 \quad (4.8)$$

An example of the fittings of the time series from Fig. 4.3 (c) by the nonlinear models NM_1 and NM_2 is shown in Fig. 4.6. Neither NM_1 nor NM_2 captured the dynamic features of the time series. The free energy criterion can be applied here to compare the models with the same order: the free energy of M_2 as shown in Table 4.1) is the largest among the three models ($\mathcal{F}_{NM_1} = -270$, $\mathcal{F}_{NM_2} = -283$ and $\mathcal{F}_{M_2} = -268$); the oscillatory dynamic behaviour of the data is not captured by NM_1 or NM_2 . The NRMSE criterion was applied here for models of different order: both the NRMSE values for the nonlinear models ($\text{NRMSE}_{NM_1} = 0.080$ for NM_1 and $\text{NRMSE}_{NM_2} = 0.067$ for NM_2) are larger than NRMSE for M_3 ($\text{NRMSE}_{M_3} = 0.053$). Additionally it is clear that the fitting using NM_2 leads to an unstable solution. Therefore, the linear model M_3 shown in (4.5) outperformed both NM_1 and NM_2 , and was chosen as the final model.

This example is a typical (representative) fitting for all the other time series in the cohort. There are a variety of nonlinear forms which could be considered for the functions f_i in (4.1a). However, with the sparse measurement time series available, a simple form is preferred, and that is why nonlinearity is introduced to the second order model before settling on the choice of the third order linear model M_3 . However, neither of the chosen forms of the nonlinear models showed better performance than the linear model M_3 . Therefore, M_3 was finally chosen as the best model.

4.4.2 Model validation

So far, the third order linear model M_3 has been chosen as the best model. An empirical validation method has been applied to check that M_3 did not overfit the data compared with M_2 by including one more extra degree of freedom in the parameter space. If a model overfitted the data, the predictive ability of the model deteriorates. The validation method, known as the leave-one-out cross validation technique, has been applied to compare between the second order model M_2 and the third order model M_3 [200]. For

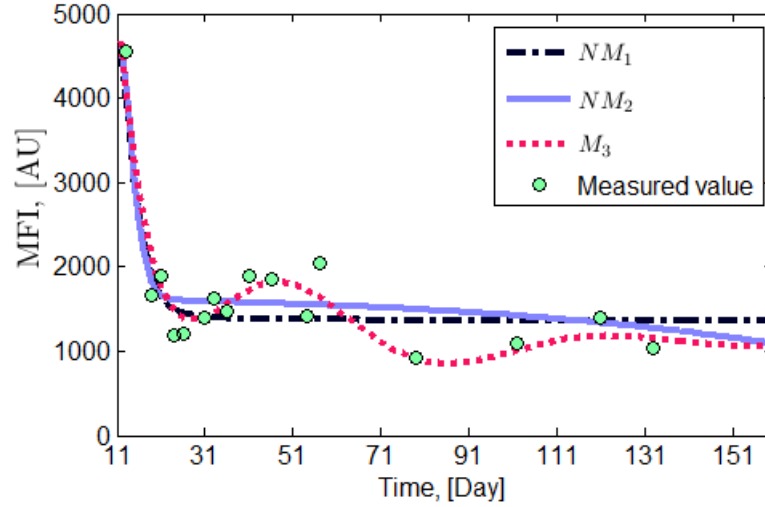


FIGURE 4.6: Fitting results compared between the two nonlinear models NM_1 and NM_2 and the linear model M_3 for the time series shown in Fig. 4.3 (c). The measured values are indicated by circles.

each time series in the cohort, we performed the following procedure for both models M_2 and M_3 :

- 1) Leave data point i out from the measurement time series, fit the model M ($M = M_2$ or M_3) based on the rest of the data points using the VB method, then after the parameters are inferred, obtain an estimation for data point i by substituting the inferred deterministic parameters, and compute the estimation error term ($e_i = y_i - \hat{y}_i$, where y_i is the measurement and \hat{y}_i is the estimation);
- 2) Repeat step one for $i = 1, \dots, n$
- 3) Compute the RMSE from e_1, \dots, e_n for both models, and choose the model with the smallest error.

The differences in the RMSE between M_2 and M_3 for each time series have been calculated and they are presented in the boxplot in Fig. 4.7. It is clear that in most cases M_3 has less errors compared with M_2 between the observations and the estimations. The differences in the errors estimated by M_2 and M_3 are tested by a one sample t-test with significant level of 0.05. The test results rejected the hypothesis that the mean of the differences between M_2 and M_3 is zero, indicating that M_3 is a better model compared with M_2 .

It is worth noticing that there are cases when M_2 is the better model; however, the aim is to identify a model that is capable of describing all of the time series and M_2 is a special case of M_3 . Therefore, the choice of M_3 over M_2 was validated.

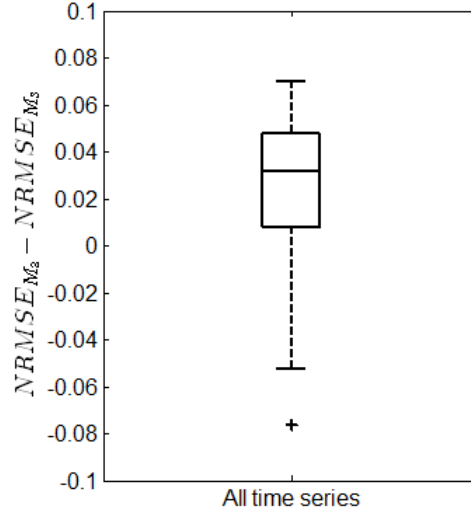


FIGURE 4.7: NRMSE value of the errors between the observations and the estimated values by M_2 and M_3

4.4.3 Structural identifiability and parameter sensitivity

A structural identifiability analysis as introduced in Section 2.6 has been performed for M_3 . For the third order linear system

$$\ddot{x}_t + \theta_3 \ddot{x}_t + \theta_2 \dot{x}_t + \theta_1 x_t + \theta_0 = 0 \quad (4.9)$$

with observations

$$y_t = x_t \quad (4.10)$$

and initial conditions

$$\mathbf{x}_0 = \begin{pmatrix} x_0 \\ \dot{x}_0 \\ \ddot{x}_0 \end{pmatrix} \quad (4.11)$$

The Laplace transform of (4.9) is as follows:

$$(s^3 X - s^2 x_0 - s \dot{x}_0 - \ddot{x}_0) + \theta_3(s^2 X - s x_0 - \dot{x}_0) + \theta_2(s X - x_0) + \theta_1 X + \theta_0 s^{-1} = 0 \quad (4.12)$$

Rearranging (4.12), the following form can be obtained:

$$X = \frac{(x_0 + \dot{x}_0 + \ddot{x}_0)s^3 + (\theta_3 x_0 + \dot{x}_0)s^2 + (\theta_2 x_0 + \theta_3 \dot{x}_0 + \ddot{x}_0)s - \theta_0}{s^4 + \theta_3 s^3 + \theta_2 s^2 + \theta_1 s} \quad (4.13)$$

where θ_3 , θ_2 , θ_1 , $x_0 + \dot{x}_0 + \ddot{x}_0$, $\theta_3 x_0 + \dot{x}_0$, $\theta_2 x_0 + \theta_3 \dot{x}_0 + \ddot{x}_0$ and θ_0 are assumed to be known [93]. Therefore, θ_3 , θ_2 , θ_1 and θ_0 are uniquely identifiable.

To check parameter sensitivity, the procedures explained in Section 2.7 – the one-at-a-time parameter sensitivity analysis – were performed for model M_3 (see details in Section 2.7). To check how sensitive the output is to a small change in each deterministic parameter, the following steps have been performed for model M_3 :

- 1) Simulate the time series with no measurement or system noise and obtain the root mean square error (RMSE) between the noise free time series and the measurement time series.
- 2) Take 1000 random samples of $\theta_i^{(j)}$ ($j = 1, 2, \dots, 1000$) from a uniform distribution from $0.99\hat{\theta}_i$ to $1.01\hat{\theta}_i$, where $\hat{\theta}_i$ is the posterior mean of θ_i .
- 3) Calculate the RMSE values between the measurement values and each of 1000 generated time series, denoted as $\text{RMSE}^{(j)}$, using the sampled parameter $\theta_i^{(j)}$ and the posterior means of the rest of the parameters.
- 4) Using (2.83), obtain the sensitivity index SI for parameter θ_i .

The same procedure can be performed to the inferred parameters of all the time series. Considering the parameter values for different time series fitted by M_3 are in the same neighbourhood, the sensitivity indices SI of θ_1 , θ_2 , θ_3 , and θ_0 in M_3 for one typical time series as shown in Fig. 4.6 is presented as an example here. As shown in Table 4.2, 1% change in θ_1 in M_3 can cause 0.17% change in RMSE. The RMSE values between the measurements and the output, generated when the posterior means of the parameters were used, is $\text{RMSE}^{(j)} = 239.0$ AU. With 1% perturbation in $\hat{\theta}_1$, $\text{RMSE}^{(j)}$ is between 239.3 AU to 240.2 AU. According to Table 4.2, the RMSE values remain within a small range when each of the four parameters are perturbed within 1%, the model M_3 is robust around the inferred posterior means of the parameters.

TABLE 4.2: Summary of the parameter sensitivities for M_3

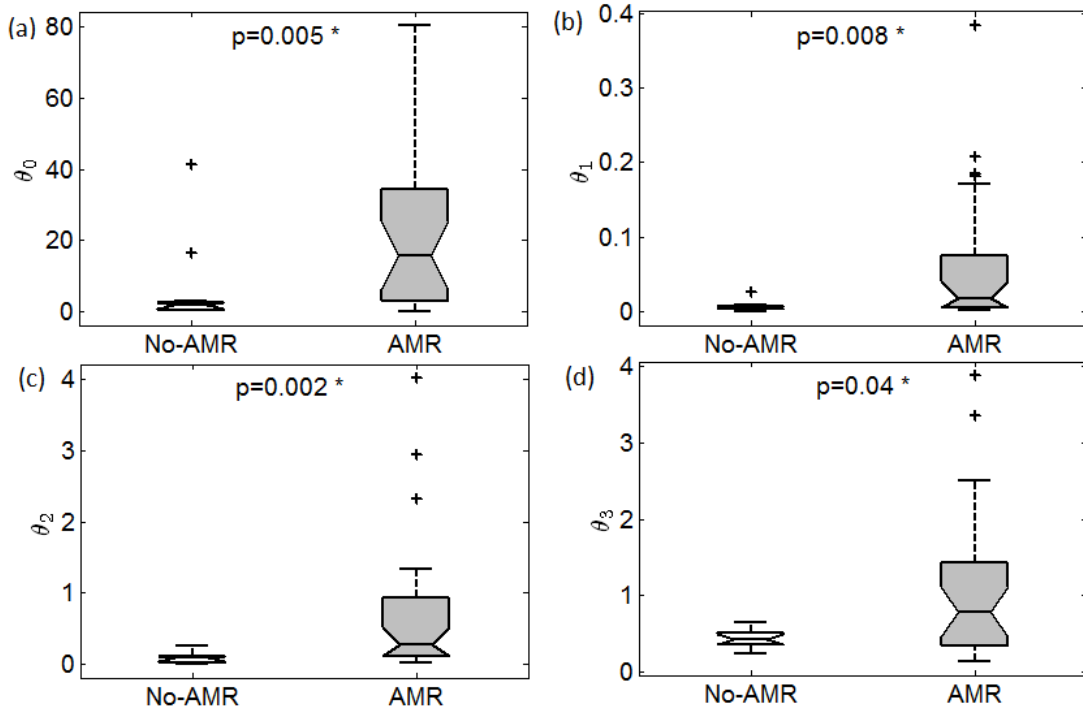
Parameter	SI	RMSE ^(j) range
θ_1	0.17	239.3 – 240.2 AU
θ_2	0.11	239.4 – 240.1 AU
θ_3	0.30	239.4 – 241.5 AU
θ_0	-0.19	239.2 – 240.1 AU

4.4.4 Analysis of the inferred parameters

Statistical analysis of the model parameters was performed using the Wilcoxon rank sum test. The null hypothesis of no difference between the groups of interest was tested at the 5% significance level, and the results are presented as p -values.

4.4.4.1 Comparison of deterministic parameters

The inferred parameters (θ_0 , θ_1 , θ_2 and θ_3) of the selected model M_3 have been compared between the two groups (AMR and no-AMR) for meaningful differences. The results are presented in Fig. 4.8 (a) – (d) in the form of boxplots. For all four parameters,

FIGURE 4.8: Boxplot for the inferred parameters θ_0 , θ_1 , θ_2 , θ_3

the ranges of the parameter values are much wider in the AMR group compared with

the no-AMR group, indicating more diverse dynamic behaviour of DSA in the AMR group. The Wilcoxon rank sum test showed statistically significant differences in the median values between the AMR and the no-AMR group for all four parameters, which confirmed the results of our preliminary study [157] with fewer cases.

Even though the values of the parameters do not have direct clinical interpretations, which is one of the main drawbacks of data-driven modelling in biomedical research, a certain combination of the parameters indicates important features of the system under investigation. The ratio θ_0/θ_1 from (4.5) defines the settling level of DSA, which is of clinical interest. Kidney transplantation constitutes a major disturbance in the immune system, and the system should settle down to a new homeostatic equilibrium after the transient response to the transplanted organ. A successful transplantation is usually characterised by a new stable steady state with low DSA levels (ideally zero, or below the threshold of the detection assay). From the comparison between the AMR and no-AMR groups shown in Fig. 4.9, the majority of the settling MFI values in both groups are less than 1000 AU, indicating low DSA settling levels. The highest settling level in the no-AMR group is 3862 AU, compared with the level of 5783 AU in the AMR group. The lowest settling level in the no-AMR group is 22 AU, compared with the level of 27 AU in the AMR group. As shown in Fig. 4.9, there is no significant difference with a

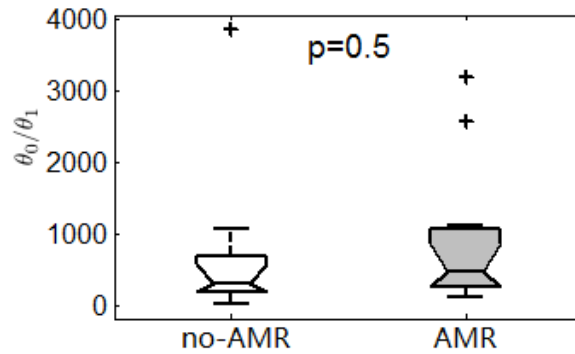


FIGURE 4.9: Boxplot for the settling values compared between no-AMR and AMR groups

p -value of 0.5 in the median value of θ_0/θ_1 between the groups (300 AU in the no-AMR group, 425 AU in the AMR group), which means that a DSA time series from the AMR group does not necessarily have a higher settling level. However, significant difference in θ_0 and θ_1 separately between the groups shown in Fig. 4.8 implies that the dynamic behaviour of DSA in the AMR group might be controlled by more complex and diverse underlying mechanisms.

Such a detailed analysis of the parameters of the models developed allows for enhanced understanding of the clinical characteristics which are most important for successful outcome in this high risk form of transplantation. Our findings may facilitate the formation of an accurate pre-transplant risk profile which predicts AMR and allows the clinician to intervene at a much earlier stage. Given that AMR in the early post-transplant period has been shown to lead to worse long-term graft outcomes any strategy to prevent early AMR will be of great benefit to the patients [191].

4.4.4.2 Comparison of stochastic parameters

Noise accounts for both measurement error due to inaccuracy in the MFI readings, and the perpetual actions of many unaccounted for factors that influence the evolution of the system. The noise intensities I_ϵ were compared between the no-AMR and AMR groups. In our preliminary analysis [157] of a limited number of time series (9 for the no-AMR group and 12 for the AMR group), the no-AMR group had a smaller and more compact range of the noise intensities compared with the AMR group (shown in Fig. 5 of [157]). Limited by the numbers of cases available, the Wilcoxon rank sum test showed no significant difference in the median value between groups with a p -value of 0.08. This study, on a larger cohort with almost twice as many time series, confirmed the previous observation with a smaller p -value of 0.01, indicating a significant difference in the median values of the noise intensities between groups.

The square root of the noise intensity $\sqrt{I_\epsilon}$, which is an absolute error value, shares the same unit as the MFI level. In the no-AMR group with an average MFI peak height of 5716 AU, the median (and range) for $\sqrt{I_\epsilon}$ were 159 AU (5 AU – 353 AU). In the AMR group with an average MFI peak height of 8502 AU, the median (and range) for $\sqrt{I_\epsilon}$ were 253 AU (34 AU – 1425 AU). A smaller noise intensity and more compact range of values across the no-AMR group is noticeable. Even though the assay used to measure the DSA level in both groups was the same, the relationship between MFI measurement and the antibody level deviates from linearity as the antibody level approaches 10,000 AU. The higher antibody peak values in the AMR group can therefore introduce an additional source of measurement error compared with the no-AMR group, explaining the wider range and greater magnitude of the noise intensity seen in the AMR group. Another explanation could be a different level of model imperfection between the two

groups. The higher level of noise in the AMR group could be caused by more and/or stronger unconsidered factors in M_3 . Also it is worth noticing that the priors for the noise intensity applied in the inference method are chosen to be weakly informative. A more informative prior may limit the flexibility of the model, but a carefully chosen informative prior could improve the estimation of deterministic parameters and parameters related to the noise description. The choice of the priors is not straightforward, and an appropriate methodology of choosing the best priors will be further developed in the future carried on by a new PhD project described in Chapter 5.

In single antigen bead measurements, another measure, termed the *inter-assay coefficient of variability* (CV) is often used to indicate the measurement uncertainty. It is defined as the ratio of the standard deviation and the mean value of several measurements using separate assays. In [201], the inter-assay CV was larger than 20% when the measurements from seven different labs were compared. In our model, considering the median value of $\sqrt{I_\varepsilon}$ and the median value of MFI measurements, the median CV is 13% and 14% for the no-AMR and AMR groups respectively, which is less than the 20% given in [201].

4.4.5 Eigenvalues

The evolution equation (4.5) can be transformed into the third order linear state space model of the form

$$\begin{pmatrix} \dot{x}_t \\ \ddot{x}_t \\ \dddot{x}_t \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\theta_1 & -\theta_2 & -\theta_3 \end{pmatrix} \begin{pmatrix} x_t \\ \dot{x}_t \\ \ddot{x}_t \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \theta_0 \end{pmatrix} \quad (4.14)$$

The solution of (4.14) is defined by the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of the 3×3 matrix, the corresponding eigenvectors and three initial conditions. The sum of the eigenvalues defines the divergence of the vector field (phase volume $V(t)$) in the state space [202]:

$$V(t) = V_0 e^{(\lambda_1 + \lambda_2 + \lambda_3)t} = V_0 e^{Rt}, \quad (4.15)$$

where R can be interpreted as the dissipation rate of DSA. For all of the time series in the cohort, the dissipation rate is less than zero, which means that the phase volume

shrinks. Analytically, R equals the trace of this 3×3 matrix:

$$R = -\theta_3 \quad (4.16)$$

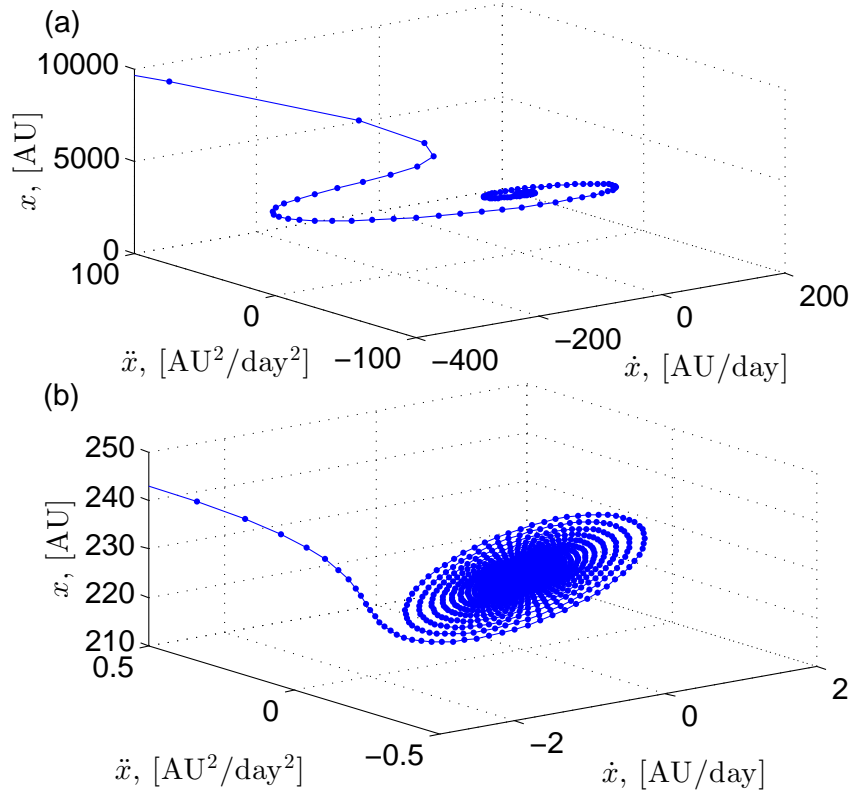


FIGURE 4.10: Phase portraits of the three dimensional system for two DSA time series, (a) from a patient in the AMR group and (b) from a patient in the no-AMR group. The time difference between two consecutive markers is one day.

The eigenvalues for every DSA time series were calculated using the inferred parameters θ_1 , θ_2 and θ_3 . Each DSA time series in the cohort is characterised by three eigenvalues, one of which is real, λ_1 , and two of which are complex conjugate, $\lambda_{2,3} = \lambda_r \pm i\lambda_i$. Generally speaking, all of the eigenvalues can be real, but one real eigenvalue and two complex eigenvalues are found for all the DSAs in the cohort. All eigenvalues λ_1 and the real parts of λ_2 and λ_3 were negative, confirming that the system generates stable solutions for each DSA type, which satisfies criterion 4 in Section 4.3.3. The system dynamics for each DSA demonstrate a decay with some oscillations, the frequency of which is determined by λ_i . There was no significant difference ($p > 0.05$ as in Fig. 4.11 (a) (b)) in either the characteristic times, associated with the largest or smallest real parts of the eigenvalues, for the AMR and no-AMR groups. The characteristic dissipation rate R takes into account the overall decay along the path from the peak

value down to the steady state. To visualise the dynamics of DSA, phase portraits have been plotted for an AMR case (Fig. 4.10 (a)), and a no-AMR case (Fig. 4.10 (b)). The trajectories start from the inferred initial states and evolve to the fixed points in a spiral manner in the phase space. It can be seen that the dissipation rate in the AMR group (Fig. 4.10 (a), $R_{(a)} = -0.81 \text{ days}^{-1}$) is faster than the no-AMR group (Fig. 4.10 (b), $R_{(b)} = -0.27 \text{ days}^{-1}$).

The dissipation rates and frequencies of oscillations were compared between the AMR and no-AMR groups for the 39 time series. In the no-AMR group, the median and range (in brackets) for R were -0.42 (-0.66 — -0.25) days^{-1} . In the AMR group, the median and range (in brackets) for R were -0.79 (-3.88 — -0.15) days^{-1} . The comparison of the dissipation rates R between the groups for all of the time series confirmed a significantly faster dissipation rate of DSA in the AMR group than in the no-AMR group with a p -value of 0.04 (shown in Fig. 4.11 (c)).

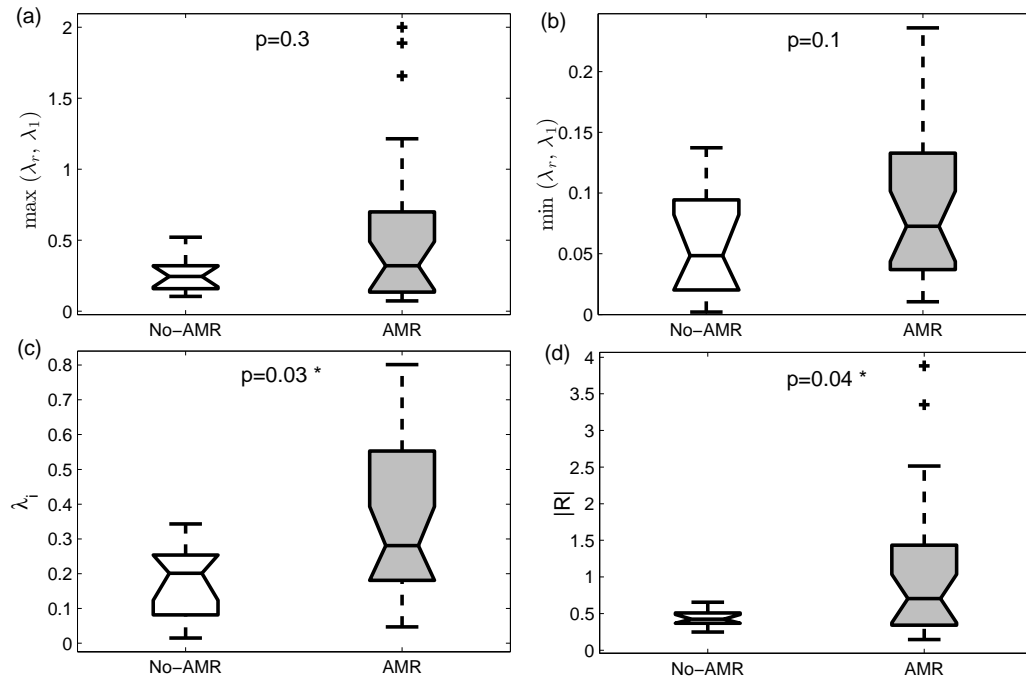


FIGURE 4.11: Boxplot for (a) the larger real part of the eigenvalues; (b) the smaller real part of the eigenvalues; (c) the imaginary part of the eigenvalues $\lambda_{2,3}$; (d) the dissipation rate between the two groups.

The imaginary parts of the eigenvalues between the AMR and no-AMR groups also showed significant differences with a p -value of 0.03 (shown in Fig. 4.11 (d)). In the no-AMR group, the median and range (in brackets) for λ_i were 0.20 (0.01 — 0.34) days^{-1} .

In the AMR group, the median and range (in brackets) for λ_i were 0.28 (0.05 — 0.80) days⁻¹. The larger values of the imaginary parts in the AMR group represent a higher frequency of oscillation, which indicates a stronger regulation during the transient antibody response for the patients in the AMR group. One hypothesis for this regulation is the possible production of a secondary antibody (such as anti-idiotypic) which targets the dramatically increased DSA, resulting in a battling force between the DSA production and secondary antibody production [193].

Note that the previous study by Higgins et al. [156] investigated the change in absolute MFI values and in the mean percentage falls in the AMR and no-AMR groups, and suggested that the falls were greater in the AMR group compared with the no-AMR group. Our results show that not only is the difference in the MFI level between peak and steady state different between the two groups, but the rate of change of the fall is faster in the AMR group, also implying a stronger regulation mechanism in this group.

4.5 Conclusions

With a unique dataset of DSA time series available, a mathematical model in the form of differential equations has been developed for the first time to describe the dynamics of the ‘falls’ in DSA for patients with and without AMR episodes. A third order linear model was selected as it successfully captured the common features of the falling dynamics in DSA during the early post-transplant stage in the AMR and no-AMR groups. The model has proved useful in classification between two clinically different groups. Even though the settling level of the DSA, which can be observed from the clinical data, showed no difference between the AMR and the no-AMR groups, all the parameters of the model (both deterministic and stochastic) were found to be significantly different between the two groups. This approach is found to be useful in capturing properties of antibody evolution from their peak concentration to the final settling level and showed that the dynamic responses are different in AMR and no-AMR groups. A higher frequency of oscillations and a faster antibody dissipation rate for the AMR group has been observed from the phase portraits depicting the trajectories of the system states, and a further test confirmed significant differences between the groups.

The findings have important implications for the development of laboratory assays that might define the nature of the mechanisms responsible for the falls in DSA levels post-transplant, since a fuller understanding of these mechanisms might allow for pre-transplant manipulation of DSA levels and improved clinical outcomes. This is particularly important with respect to the oscillating nature of DSA levels, which may reflect a system slowly reaching homeostasis, and may be reflected in laboratory measurements.

Further work might also include modelling in relation to more detailed characteristics of the antibodies. For example, we have already shown that the subclasses of IgG are associated with clinical outcomes, so that measuring the levels of these subclasses at more time points might be valuable [203]. The clinical outcome measures might also be extended. Since acute antibody-mediated rejection is often treatable and is not always associated with a poor clinical outcome (especially when the settling level of DSA is very low), longer term graft survival could also be considered as an important outcome level. Day to day renal function does not always follow DSA levels [190] and our understanding of how a graft responds to DSA levels and how AMR evolves is limited.

This study presents the results on data-driven model development for early post-transplant antibody dynamics, focusing on one of the typical patterns of a rapid fall following a rapid rise in DSA after kidney transplantation. Future work will involve classification and modelling of the other patterns of the post-transplant DSA dynamics that have been described in Section 4.2 and the development of a universal model that is capable of describing such non-typical dynamic patterns in DSA evolution.

Chapter 5

Conclusions and future work

This thesis developed and presented data-driven model identification techniques, consisting of selecting and developing models based on one-dimensional clinical dynamic data in the field of biomedical engineering. The dynamic data contain essential information about the evolution of the underlying physiological system; dealing with these data is challenging due to the limited accessible measurements. Targeting at this practical difficulty, a data-driven model development strategy has been devised to extract information carried by a measurement time series of a single variable about the underlying complex system. The original project aims and objectives have been fully met. The developed models are chosen to be based on nonlinear stochastic differential equations, which was selected due to the nonlinear, stochastic and continuous nature of the underlying physiological system. Recognising the nonlinearity as one of the fundamental system characteristics, we hypothesised and confirmed the form of the dynamic equations to express various degrees of nonlinearity by different orders of polynomials. To incorporate stochasticity into the model framework, system noise and measurement noise have been accounted for separately in the system equation and measurement equation. Being data-driven, the complexity of the model is highly dependent on the quality and availability of the measurements. The model development strategy allowed selection of the best model with the appropriate degree of complexity based on the measurements available.

With the principle of parsimony at its heart, the variational Bayesian method was applied to select the model structure with the appropriate level of complexity and to infer

the model parameters. The VB method has been known since the 1990s, but its efficiency in parameter estimation and model selection has not been fully recognised outside of the area of neural-imaging for which it has been developed. It is hoped that this thesis can bring an awareness of the VB method to the biomedical engineering community, by stressing its advantages compared with other approaches, such as the maximal likelihood method, maximum-a-posteriori method, etc. Compared with sampling methods such as Markov Chain Monte Carlo and Sequential Monte Carlo methods, the VB method is deterministic and therefore much more efficient for performing Bayesian inference. It has been widely applied to various models in the literature, such as Hidden Markov Chain models [204], mixed effect models [205] etc. The VB method maximises the value of free energy by optimising the parameters and the states of the system. The free energy contains an intrinsic penalisation for model complexity, and does not require an additional penalty term to account for model complexity, unlike other model selection criteria such as the AIC and the BIC. However, the heavy penalisation for model complexity by the free energy can cause a bias towards choosing over-simplified models, especially when the measurements are too sparse to support complex model structures. Sparse data are one of the common problems in biomedical research, and all parameter inference approaches including the VB method have difficulties. However, the successful application of the VB approach in Chapter 4 showed its potential in dealing with sparse data. For the task of model selection, other criteria, such as the stability of the model, minimisation of the root mean square error, have been considered for each application, alongside the free energy criterion, to determine the most suitable parsimonious model as shown in Chapter 3 and Chapter 4.

In the first application (Chapter 3), a novel generalised model was successfully constructed to describe each of 132 postprandial glucose excursions from fifteen subjects with and without DM. The postprandial glucose excursion usually lasts for several hours before the glucose concentration settles to a steady state. Monitoring and controlling the glucose variability during this period is important for people with DM to achieve the goal of maintaining *glycaemic stability* in daily life. With the glucose measurements taken by the CGM device every five minutes, the amount of data was sufficient to support the complexity level of the second order nonlinear stochastic model. The VB method was successfully applied to select the order of the polynomial coefficients accounting for system nonlinearity. Based on the free energy criterion and the decay

ratio criterion which was devised to keep the glucose oscillations realistic, a second order nonlinear stochastic model with second order polynomials in the damping term was selected as the generalised model. Structural identifiability and parameter sensitivity analysis were applied to confirm the reliability of the model. A successful comparison with a well known physiologically based ‘*maximal model*’ taken from literature was performed. Three time series were generated by the maximal model with 35 parameters for a control and two pre-DM subjects, and were successfully fitted by the developed model with 2 deterministic parameters. It cannot be denied that the parameters in the maximal model have direct physiological interpretation, however, the successful fitting of the simulated time series using the maximal model by the developed model clearly demonstrated the parsimony of the developed model.

Comparing between the control group and the DM group, most time series in the control group (subjects without DM) could be described sufficiently well by a second order linear model with a constant damping term, which is a particular case of the nonlinear model. However, the nonlinear model with a second order polynomial damping term was selected for most glucose excursions from the group of people with DM. As an initial visual observation of the DM profiles can detect and suspect nonlinear dynamics in the DM profiles, the developed nonlinear model confirmed such observations and successfully adapted a parsimonious form to capture such nonlinearity. Such a difference in the model selection between the groups can be explained from the physiological point of view: the impaired glucose regulation for people with DM may drive the glucose variations to exhibit more nonlinear behaviour, reflected by the preference of the nonlinear model for the DM group.

Investigation of the clinical relevance of the developed model revealed some useful properties. The inferred parameter values can serve as classifiers for individual profiles of the control group and the DM group, because the deterministic and the stochastic parameters for the control and the DM group were found to be in different ranges. The simulated pre-DM subjects were fitted by the linear model, and the parameter values corresponding to these pre-DM subjects were found to be located between the parameter ranges of the controls and the DMs. This implies a clinical potential of the developed model to diagnose pre-DMs through measurement of the dynamic blood glucose excursions after food intake. The food impact parameter values for a newly diagnosed

DM patient without any treatment were significantly larger than for the other DM patients who took regular medication. This might reflect the effectiveness of medication intervention given to the DM patients.

The developed glucose dynamic model can be further extended by considering more factors that were not included in the model. The developed model only takes a single input — the food intake — into account. Other factors, such as physical exercise and mental stress, are known to impact on the glucose regulation system, and can be included as multiple inputs or additional dynamic variables in the model. In this work, food intake was considered as an impulsive excitation force; however, glucose absorption into the system does not happen instantaneously, and therefore a function of time can be explored as the input function to reflect the real event. With respect to the stochastic terms, only additive noise has been considered in this thesis. However, in the postprandial glucose application, the CGM devices usually have a larger measurement noise when the readings of the glucose concentration is high. Therefore, a different measurement equation with multiplicative measurement noise could be considered.

In the second application (Chapter 4), a third order linear model was successfully developed for the first time to describe the ‘falling’ dynamic patterns exhibited by 39 post-transplant antibody time series from 23 patients who underwent a high risk antibody incompatible kidney transplantation (AiT). These measurements were taken by a team of world-leading doctors and clinical researchers at University Hospitals Coventry and Warwickshire (UHCW), which is the leading clinical centre in the UK for antibody incompatible kidney transplantation. This clinical centre proposed a National Registry, which is the first of its type in the world and has hosted more than one hundred complex cases since 2003. The post-transplant antibody dynamics manifest a variety of patterns, among which a pattern of a clear rise and fall is shared by 39 post-transplant antibody time series that belong to two groups of patients with and without episodes of acute AMR.

The VB method was successfully applied to select the best model and infer its parameters. The choice between the stochastic model and the deterministic model, and between the linear model and the nonlinear model was made by comparing the free energy values for each time series. The normalised root mean square error criterion, accounting for

the peak height of the time series, was devised to select the order of the system. Overcoming the difficulty with the sparseness of the data and the irregularity of the sampling frequency, the third order linear model was selected and confirmed by the leave-one-out cross validation technique. The VB method showed a robust performance of parameter inference with limited numbers of measurement data points. The combined consideration of the free energy criterion and the normalised root mean square error criterion for model selection was proved to be a good strategy when data are sparse. Structural identifiability and sensitivity analysis have also been applied to confirm the reliability of the model. The developed model is structurally identifiable and is robust around the inferred parameter values.

The ranges of the parameter values were found to be significantly different between the AMR and no-AMR group. The post-transplant antibody dynamics was found to have a significantly higher frequency of oscillations and a faster antibody dissipation rate for the AMR group. These findings have important implications for the development of laboratory assays that might define the nature of the mechanisms responsible for the falls in the post-transplant antibody levels.

This work can be further extended by considering more varieties of dynamic patterns that manifest in the post-transplant antibody time series. Only time series with a dramatic rise and fall after transplantation have been modelled. Other dynamic features, such as an unsettled oscillations and a slow rise after the fall, require further investigation. Different forms of nonlinearity or different model orders might be considered to reflect the diverse dynamic patterns. In addition, the measurement noise was modelled to have a constant noise intensity regardless of the measured value. However, the linear relationship between the antibody concentration and the MFI readings (measurements) breaks when $\text{MFI} < 1000 \text{ AU}$ or $> 10,000 \text{ AU}$. This suggests that a non-additive noise can be considered.

Overall, the nonlinear model identification techniques developed in this thesis are limited to a polynomial form of nonlinearity. Such techniques can be extended to incorporate other forms of nonlinearity, such as exponential, logarithmic, trigonometric functions, into the model development process. However, in the context of biomedical system identification, the dynamic patterns exhibited in the measured time series should direct the incorporation of different nonlinear forms.

This dissertation forms a solid foundation for future research to account for the factors that have not been included into the models. Two Ph.D students have been recruited to continue the research initiated by this Ph.D project and further develop and validate the models. The two projects are:

- 1) Personalised predictive modelling and control of blood glucose dynamics.
- 2) Mathematical modelling for clinical outcome prediction in kidney transplantation.

Bibliography

- [1] M. Ghassemi, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, and M. Feng, “A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data,” *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 446–453, 2015.
- [2] S. Eslami, A. Abu-Hanna, E. de Jonge, and N. F. de Keizer, “Tight glycemic control and computerized decision-support systems: a systematic review,” *Intensive Care Medicine*, vol. 35, pp. 1505–1517, 2009.
- [3] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, “Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success,” *British Medical Association*, vol. 330, p. 765, Apr. 2005.
- [4] E. W. Steyerberg, *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media, 2008.
- [5] D. Collett, *Modelling survival data in medical research*. CRC press, 2015.
- [6] L. B. Sheiner, B. Rosenberg, and V. V. Marathe, “Estimation of population characteristics of pharmacokinetic parameters from routine clinical data,” *Journal of Pharmacokinetics and Biopharmaceutics*, vol. 5, pp. 445–479, July 1977.
- [7] N. Tsamandouras, A. Rostami-Hodjegan, and L. Aarons, “Combining the ‘bottom up’ and ‘top down’ approaches in pharmacokinetic modelling: fitting PBPK models to observed clinical data,” *British journal of clinical pharmacology*, vol. 79, pp. 48–55, Jan. 2013.
- [8] V. Díaz, M. Viceconti, K. Stroetmann, and D. Kalra, “Roadmap for the digital patient,” *European Commission*, 2013.

- [9] J. P. Sturmborg and C. M. Martin, *Handbook of systems and complexity in health*. 2013.
- [10] J. Paduart, L. Lauwers, J. Swevers, K. Smolders, J. Schoukens, and R. Pintelon, “Identification of nonlinear systems using polynomial nonlinear state space models,” *Automatica*, vol. 46, no. 4, pp. 647–656, 2010.
- [11] C. W. Chan, “Editorial: special issue on data-driven modelling methods and their applications,” *International Journal of Systems Science*, vol. 34, pp. 731–732, Nov. 2003.
- [12] L. Dewasme, P. Bogaerts, and A. V. Wouwer, “Monitoring of Bioprocesses : Mechanistic and Data-Driven Approaches,” in *Monitoring of Bioprocesses: Mechanistic and Data-Driven Approaches*, ch. 3, pp. 57–97, 2009.
- [13] J. Schoukens and R. Pintelon, *Identification of linear systems: a practical guideline to accurate modeling*. Elsevier, 2014.
- [14] D. Rickles, P. Hawe, and A. Shiell, “A simple guide to chaos and complexity,” *Journal of epidemiology and community health*, vol. 61, no. 11, pp. 933–937, 2007.
- [15] J. J. Batzel, M. Bachar, and F. Kappel, *Mathematical modeling and validation in physiology: applications to the cardiovascular and respiratory systems*, vol. 2064. Springer, 2012.
- [16] A. Gelman, F. Bois, and J. Jiang, “Physiological pharmacokinetic analysis using population modelling and informative prior distributions,” *Journal of the American Statistical Association*, vol. 91, no. 439, pp. 1400–1412, 1996.
- [17] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*. Cambridge University Press, 2003.
- [18] S. M. Pincus and A. L. Goldberger, “Physiological time-series analysis: what does regularity quantify,” *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 266, no. 4, pp. H1643–H1656, 1994.
- [19] A. Guerrini, *Experimenting with humans and animals: from Galen to animal rights*. JHU Press, 2003.

- [20] H. B. Van Der Worp, D. W. Howells, E. S. Sena, M. J. Porritt, S. Rewell, O. Collins, and M. R. Macleod, “Can animal models of disease reliably inform human studies,” *PLoS Medicine*, vol. 7, no. 3, p. e1000245, 2010.
- [21] J. G. Chase, A. J. Le Compte, J.-C. Preiser, G. M. Shaw, S. Penning, and T. Desai, “Physiological modeling, tight glycemic control, and the ICU clinician: what are models and how can they affect practice,” *Annals of Intensive Care*, vol. 1, no. 1, pp. 1–8, 2011.
- [22] V. Z. Marmarelis, *Nonlinear dynamic modeling of physiological systems*. John Wiley & Sons, 2004.
- [23] O. Nelles, *Nonlinear system identification: from classical approaches to neural networks and fuzzy models*. Springer Science & Business Media, 2013.
- [24] H. Garnier and L. Wang, *Identification of Continuous-time Models from Sampled Data*. Advances in Industrial Control, London: Springer London, 2008.
- [25] A. R. Bergstrom and W. Park, “Optimal forecasting of discrete stock and flow data generated by a higher order continuous time system,” *Computers & Mathematics with Applications*, vol. 17, no. 8, pp. 1203–1214, 1989.
- [26] S. M. Dunn, A. Constantinides, and P. V. Moghe, “Dynamic Systems : Ordinary Differential Equations,” in *Numerical Methods in Biomedical Engineering*, ch. 7, pp. 209–287, 2007.
- [27] L. Ljung and T. Glad, *Modeling of dynamic systems*. Englewood Cliffs: PTR Prentice Hall, 1994.
- [28] E. Sejdi and L. A. Lipsitz, “Necessity of noise in physiology and medicine,” *Computer methods and programs in biomedicine*, vol. 111, no. 2, pp. 459–470, 2013.
- [29] E. Bloch-Salisbury, P. Indic, F. Bednarek, and D. Paydarfar, “Stabilizing immature breathing patterns of preterm infants using stochastic mechanosensory stimulation,” *Journal of applied physiology*, vol. 107, pp. 1017–1027, Oct. 2009.
- [30] A. A. Priplata, J. B. Niemi, J. D. Harry, L. A. Lipsitz, and J. J. Collins, “Vibrating insoles and balance control in elderly people,” *Lancet*, vol. 362, pp. 1123–1124, Oct. 2003.

- [31] L. A. Lipsitz, “Physiological complexity, aging, and the path to frailty,” *Science’s SAGE KE*, vol. 2004, no. 16, p. pe16, 2004.
- [32] C.-K. Peng, S. Havlin, J. M. Hausdorff, J. E. Mietus, H. E. Stanley, and A. L. Goldberger, “Fractal mechanisms and heart rate dynamics long-range correlations and their breakdown with disease,” *Journal of Electrocardiology*, vol. 28, pp. 59–65, 1995.
- [33] A. Beuter, L. Glass, M. C. Mackey, and M. S. Titcombe, *Nonlinear dynamics in physiology and medicine*, vol. 25. Springer Science & Business Media, 2013.
- [34] D. J. Wilkinson, “Stochastic modelling for quantitative description of heterogeneous biological systems,” *Nature reviews. Genetics*, vol. 10, pp. 122–133, Feb. 2009.
- [35] K. J. Friston, “Variational filtering,” *NeuroImage*, vol. 41, no. 3, pp. 747–766, 2008.
- [36] A. Gelman and D. B. Rubin, “Markov Chain Monte Carlo Methods in biostatistics,” *Statistical Methods in Medical Research*, vol. 5, no. 4, pp. 339–355, 1996.
- [37] B. P. Carlin and T. A. Louis, “Bayes and empirical Bayes methods for data analysis,” *Statistics and Computing*, vol. 7, no. 2, pp. 153–154, 1997.
- [38] F. J. Samaniego, *A comparison of the Bayesian and frequentist approaches to estimation*. Springer Science & Business Media, 2010.
- [39] M. J. Bayarri and J. O. Berger, “The interplay of bayesian and frequentist analysis,” *Statistical Science*, vol. 19, pp. 58–80, Feb. 2004.
- [40] M. Clyde and E. I. George, “Model Uncertainty,” *Statistical Science*, vol. 19, pp. 81–94, Feb. 2004.
- [41] M. J. Beal, *Variational algorithms for approximate bayesian inference*. PhD thesis, 2003.
- [42] V. Smidl and A. Quinn, *The variational Bayes method in signal processing*. Springer Science & Business Media, 2006.
- [43] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, vol. 2. Taylor & Francis, 2014.

- [44] W. K. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, vol. 57, no. 1, p. 97, 1970.
- [45] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, vol. 39. Chapman & Hall/CRC, 1996.
- [46] B. P. Carlin and T. A. Louis, *Bayesian methods for data analysis*. CRC Press, 2008.
- [47] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2002.
- [48] D. J. C. Mackay, *Information Theory, Inference, and Learning Algorithms*. Cambridge university press, 2003.
- [49] B. T. Kulakowski, J. F. Gardner, and J. L. Shearer, *Dynamic Modeling and Control of Engineering System*. Cambridge University Press, 2007.
- [50] W. F. Trench, *Elementary Differential Equations with Boundary Value Problems*. 2013.
- [51] A. Longtin, "Effects of Noise on Nonlinear Dynamics," *Nonlinear Dynamics in Physiology and Medicine*, vol. 25, pp. 149–189, 2003.
- [52] C. Gardiner, *Stochastic methods: A Handbook for the Natural and Social Sciences*. 4 ed., 2009.
- [53] P. M. Bentler and D. G. Bonett, "Significance tests and goodness of fit in the analysis of covariance structures," *Psychological Bulletin*, vol. 88, no. 3, pp. 588–606, 1980.
- [54] C. T. H. Baker, G. A. Bocharov, J. M. Ford, P. M. Lumb, S. J. Norton, C. a. H. Paul, T. Junt, P. Krebs, and B. Ludewig, "Computational approaches to parameter estimation and model selection in immunology," *Journal of Computational and Applied Mathematics*, vol. 184, no. 1, pp. 50–76, 2005.
- [55] S. Geisser, *Predictive inference*, vol. 55. CRC press, 1993.

- [56] N. R. Kristensen, H. Madsen, and S. H. Ingwersen, "Using stochastic differential equations for PK/PD model development," *Journal of Pharmacokinetics and Pharmacodynamics*, vol. 32, no. 1, pp. 109–141, 2005.
- [57] E. T. Jaynes, *Bayesian methods: general background*. 1986.
- [58] R. M. Neal, "Probabilistic inference using Markov Chain Monte Carlo methods," Tech. Rep. September, 1993.
- [59] B. Calderhead and M. Girolami, "Estimating Bayes factors via thermodynamic integration and population MCMC," *Computational Statistics and Data Analysis*, vol. 53, no. 12, pp. 4028–4045, 2009.
- [60] J. S. Liu and R. Chen, "Sequential Monte Carlo Methods for Dynamic Systems," *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 1032–1044, 1998.
- [61] A. Doucet, N. de Freitas, and N. Gordon, "An Introduction to Sequential Monte Carlo Methods," *Sequential; Monte Carlo Methods in Practice*, pp. 3–14, 2001.
- [62] P. Del Moral, A. Doucet, and A. Jasra, "Sequential Monte Carlo samplers," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 68, no. 3, pp. 411–436, 2006.
- [63] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B Methodological*, vol. 39, no. 1, pp. 1–38, 1977.
- [64] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 362–369, 2001.
- [65] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 21–30, 1999.
- [66] W. R. Ashby, "Principles of the self-organizing system," *Facets of Systems Science*, vol. 6, no. 1991, pp. 521–536, 1991.

- [67] T. M. Cover, J. A. Thomas, J. Bellamy, R. L. Freeman, and J. Liebowitz, *Elements of Information Theory*. John Wiley & Sons, 1991.
- [68] C. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.
- [69] J. L. Jensen, “Sur les fonctions convexes et les inégalités entre les valeurs moyennes,” *Acta Mathematica*, vol. 30, no. 1, pp. 175–193, 1906.
- [70] C. M. Bishop, *Pattern recognition and machine learning*. Springer New York, 2006.
- [71] B. Wang and D. M. Titterton, “Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values,” *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 1–14, 2010.
- [72] J. Daunizeau, K. J. Friston, and S. J. Kiebel, “Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models,” *Physica D*, vol. 238, no. 21, pp. 2089–2118, 2009.
- [73] H. E. Rauch, C. T. Striebel, and F. Tung, “Maximum likelihood estimates of linear dynamic systems,” *AIAA journal*, vol. 3, no. 8, pp. 1445–1450, 1965.
- [74] K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny, “Variational free energy and the Laplace approximation,” *NeuroImage*, vol. 34, no. 1, pp. 220–234, 2007.
- [75] W. D. Penny, “Comparing dynamic causal models using AIC, BIC and free energy,” *NeuroImage*, vol. 59, no. 1, pp. 319–330, 2012.
- [76] M. A. Pitt and I. J. Myung, “When a good fit can be bad,” *Trends in Cognitive Sciences*, vol. 6, no. 10, pp. 421–425, 2002.
- [77] J. Daunizeau and L. Rigoux, “VBA toolbox,” 2009.
- [78] G. Claeskens and N. L. Hjort, *Model selection and model averaging*, vol. 330. Cambridge university press, 2008.
- [79] P. Stoica and Y. Sel, “A review of information criterion rules,” vol. 21, no. 4, pp. 36–47, 2004.

- [80] R. E. Kass and A. E. Raftery, "Bayes Factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.
- [81] B. L. Jones, D. S. Nagin, and K. Roeder, "A SAS procedure based on mixture models for estimating developmental trajectories," *Sociological methods & research*, vol. 29, no. 3, pp. 374–393, 2001.
- [82] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.
- [83] R. W. Katz, "On some criteria for estimating the order of a Markov chain," *Technometrics*, vol. 23, no. 3, pp. 243–249, 1981.
- [84] D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter, *The BUGS book: A practical introduction to Bayesian analysis*. CRC press, 2012.
- [85] A. Gelman, "Prior distributions for variance parameters in hierarchical models," *Bayesian analysis*, vol. 1, no. 3, pp. 515–533, 2006.
- [86] S. Van Dongen, "Prior specification in Bayesian statistics: three cautionary tales," *Journal of Theoretical Biology*, vol. 242, no. 1, pp. 90–100, 2006.
- [87] S. Clarke and R. Barron, "Jeffreys' prior is asymptotically under entropy risk least favorable," *Journal of Statistical planning and Inference*, vol. 41, no. 1, pp. 37–60, 1994.
- [88] C. P. Robert, N. Chopin, and J. Rousseau, "Harold Jeffreys's theory of probability revisited," *Statistical Science*, vol. 24, pp. 141–172, May 2009.
- [89] P. Stoica, "An introduction to identification," *Automatica*, vol. 24, no. 3, pp. 426–427, 1988.
- [90] R. Pintelon and J. Schoukens, *System identification: a frequency domain approach*. John Wiley & Sons, 2012.
- [91] W. A. Link and R. J. Barker, *Bayesian inference: with ecological applications*. Academic Press, 2009.
- [92] R. Bellman and K. J. Aström, "On structural identifiability," *Mathematical Biosciences*, vol. 7, no. 3-4, pp. 329–339, 1970.

- [93] K. R. Godfrey, "The identifiability of parameters of models used in biomedicine," *Mathematical Modelling*, vol. 7, no. 9-12, pp. 1195–1214, 1986.
- [94] M. Grewal and K. Glover, "Identifiability of linear and nonlinear dynamical systems," *IEEE Transactions on Automatic Control*, vol. 21, no. 6, pp. 833–837, 1976.
- [95] H. Pohjanpalo, "System identifiability based on the power series expansion of the solution," *Mathematical Biosciences*, vol. 41, no. 1-2, pp. 21–33, 1978.
- [96] M. J. Chappell, K. R. Godfrey, and S. Vajda, "Global identifiability of the parameters of nonlinear systems with specified inputs: A comparison of methods," *Mathematical Biosciences*, vol. 102, no. 1, pp. 41–73, 1990.
- [97] G. Margaria, E. Riccomagno, M. J. Chappell, and H. P. Wynn, "Differential algebra methods for the study of the structural identifiability of rational function state-space models in the biosciences," vol. 174, pp. 1–26, 2001.
- [98] Mathematica, *Wolfram Mathematica 9.0.0.0*. Wolfram, 2012.
- [99] K. E. Hines, T. R. Middendorf, and R. W. Aldrich, "Determination of parameter identifiability in nonlinear biophysical models: a Bayesian approach," *The Journal of general physiology*, vol. 143, pp. 401–416, Mar. 2014.
- [100] A. E. Gelfand and S. K. Sahu, "Identifiability , improper priors and Gibbs sampling for generalized linear models," *Journal of the American Statistical Association*, vol. 94, pp. 247–253, 1998.
- [101] D. V. Lindley, *Bayesian statistics: A review*. SIAM, 1972.
- [102] J. A. Jacquez and P. Greif, "Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design," *Mathematical Biosciences*, vol. 77, no. 1-2, pp. 201–227, 1985.
- [103] N. A. W. van Riel, "Dynamic modelling and analysis of biochemical networks: Mechanism-based models and model-based experiments," *Briefings in Bioinformatics*, vol. 7, no. 4, pp. 364–374, 2006.
- [104] D. M. Hamby, "A review of techniques for parameter sensitivity analysis of environmental models," *Environmental Monitoring and Assessment*, vol. 32, no. 2, pp. 135–154, 1994.

- [105] Z. Zi, K. H. Cho, M. H. Sung, X. Xia, J. Zheng, and Z. Sun, "In silico identification of the key components and steps in IFN- γ induced JAK-STAT signaling pathway," *FEBS Letters*, vol. 579, no. 5, pp. 1101–1108, 2005.
- [106] I. Swameye, T. G. Muller, J. Timmer, O. Sandra, and U. Klingmuller, "Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling," in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 1028–1033, 2003.
- [107] H. Schmidt and E. W. Jacobsen, "Linear systems approach to analysis of complex dynamic behaviours in biochemical networks," *Systems biology*, vol. 1, no. 1, pp. 149–158, 2004.
- [108] H. I. Wu, "A case study of type 2 diabetes self-management," *Biomedical engineering online*, vol. 4, pp. 1–9, Jan. 2005.
- [109] N. Khovanova, Y. Zhang, and T. A. Holt, "Generalised stochastic model for characterisation of subcutaneous glucose time series," *IEEE-EMBS International Conference on Biomedical and Health Informatics*, pp. 484–487, June 2014.
- [110] Y. Zhang, T. A. Holt, and N. Khovanova, "A data driven nonlinear stochastic model for blood glucose dynamics," *Computer Methods and Programs in Biomedicine*, vol. 125, pp. 19–25, 2015.
- [111] International Diabetes Federation, "International Diabetes Federation Annual report 2014," tech. rep., 2014.
- [112] C. Cobelli, "Diabetes: Models, signals, and control," *IEEE reviews in biomedical engineering*, vol. 2, pp. 54–96, 2009.
- [113] F. Ståhl and R. Johansson, "Diabetes mellitus modeling and short-term prediction based on blood glucose measurements," *Mathematical Biosciences*, vol. 217, no. 2, pp. 101–117, 2009.
- [114] E. S. Horton, "Defining the Role of Basal and Prandial Insulin for Optimal Glycemic Control," *Journal of the American College of Cardiology*, vol. 53, no. 5s1, pp. S21–S27, 2009.

- [115] J. E. Gerich, "The genetic basis of type 2 diabetes mellitus: impaired insulin secretion versus impaired insulin sensitivity," *Endocrine reviews*, vol. 19, no. 4, pp. 491–503, 1998.
- [116] Diabetes Prevention Program Research Group, "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin," *The New England Journal of Medicine*, vol. 346, no. 6, pp. 393–403, 2002.
- [117] E. L. Lim, K. G. Hollingsworth, B. S. Aribisala, M. J. Chen, J. C. Mathers, and R. Taylor, "Reversal of type 2 diabetes: normalisation of beta cell function in association with decreased pancreas and liver triacylglycerol," *Diabetologia*, vol. 54, no. 10, pp. 2506–2514, 2011.
- [118] G. C. Weir and S. Bonner-weir, "Five stages of evolving beta-cell dysfunction during progression to diabetes," *Diabetes*, vol. 53, no. s3, pp. S16–S21, 2004.
- [119] A. Makroglou, J. Li, and Y. Kuang, "Mathematical models and software tools for the glucose-insulin regulatory system and diabetes: an overview," *Applied Numerical Mathematics*, vol. 56, pp. 559–573, Mar. 2006.
- [120] V. Tresp, T. Briegel, and J. Moody, "Neural-network models for the blood glucose metabolism of a diabetic," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1204–1213, 1999.
- [121] J. A. Jacquez, *Compartmental analysis in biology and medicine*. JSTOR, 1985.
- [122] E. Ackerman, L. C. Gatewood, J. W. Rosevear, and G. D. Molnar, "Model studies of blood-glucose regulation," *The bulletin of mathematical biophysics*, vol. 27, no. 1, pp. 21–37, 1965.
- [123] V. W. Bolie, "Coefficients of normal blood glucose regulation," *Journal of Applied Physiology*, vol. 16, no. 5, pp. 783–788, 1961.
- [124] M. Braun and M. Golubitsky, *Differential equations and their applications*. Springer, 1983.
- [125] R. N. Bergman, "Toward physiological understanding of glucose tolerance: Minimal-model approach," *Diabetes*, vol. 38, pp. 1512–1527, Dec. 1989.

- [126] C. B. Landersdorfer and W. J. Jusko, “Pharmacokinetic/pharmacodynamic modelling in diabetes mellitus,” *Clinical Pharmacokinetics*, vol. 47, no. 7, pp. 417–448, 2008.
- [127] V. Marmarelis and G. Mitsis, *Data-driven modeling for diabetes*. Springer Berlin Heidelberg, 2014.
- [128] R. N. Bergman, L. S. Phillips, and C. Cobelli, “Physiologic evaluation of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose,” *Journal of Clinical Investigation*, vol. 68, no. 6, pp. 1456–1467, 1981.
- [129] K. M. Krudys, S. E. Kahn, and P. Vicini, “Population approaches to estimate minimal model indexes of insulin sensitivity and glucose effectiveness using full and reduced sampling schedules,” *American journal of physiology. Endocrinology and metabolism*, vol. 291, no. 4, pp. E716–E723, 2006.
- [130] I. F. Godsland, O. F. Agbaje, R. Hovorka, P. Vicini, C. Cobelli, A. Caumo, J. J. Zachwieja, A. Avogaro, K. Yarasheski, and D. M. “Undermodeling affects minimal model indexes: insights from a two-compartment model,” *American Journal of Physiology-Endocrinology And Metabolism*, vol. 276, no. 6, pp. E1171–E1193, 1999.
- [131] P. Palumbo, S. Ditlevsen, A. Bertuzzi, and A. De Gaetano, “Mathematical modeling of the glucose-insulin system: a review,” *Mathematical Biosciences*, vol. 244, no. 2, pp. 69–81, 2013.
- [132] R. N. Bergman, “Minimal model: perspective from 2005,” *Hormone Research*, vol. 64, no. s3, pp. 8–15, 2005.
- [133] A. De Gaetano and O. Arino, “Mathematical modelling of the intravenous glucose tolerance test,” *Journal of mathematical biology*, vol. 40, no. 2, pp. 136–168, 2000.
- [134] E. M. Watson, M. J. Chappell, F. Ducrozet, S. M. Poucher, and J. W. T. Yates, “A new general glucose homeostatic model using a proportional-integral-derivative controller,” *Computer methods and programs in biomedicine*, vol. 102, pp. 119–129, May 2011.

- [135] C. Dalla Man, R. A. Rizza, and C. Cobelli, "Meal simulation model of glucose-insulin system," *IEEE transactions on biomedical engineering*, vol. 54, pp. 307–310, Jan. 2006.
- [136] E. I. Georga, V. C. Protopappas, and D. I. Fotiadis, "Glucose Prediction in Type 1 and Type 2 Diabetic Patients Using Data Driven Techniques," in *Knowledge-Oriented Applications in Data Mining*, ch. 17, pp. 277–296, 2011.
- [137] D. B. Keenan, J. J. Mastrototaro, G. Voskanyan, and G. M. Steil, "Delays in minimally invasive continuous glucose monitoring devices: a review of current technology," *Journal of diabetes science and technology*, vol. 3, no. 5, pp. 1207–1214, 2009.
- [138] Z. Trajanoski and P. Wach, "Neural predictive controller for insulin delivery using the subcutaneous route," *IEEE transactions on bio-medical engineering*, vol. 45, pp. 1122–1134, 1998.
- [139] J. A. Florian and R. S. Parker, "Empirical Modeling for Glucose Control in Critical Care and Diabetes," *European Journal of Control*, vol. 11, pp. 601–616, Jan. 2005.
- [140] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, "Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series," *IEEE transactions on bio-medical engineering*, vol. 54, pp. 931–937, May 2007.
- [141] A. Gani, A. V. Gribok, S. Rajaraman, W. K. Ward, and J. Reifman, "Predicting Subcutaneous Glucose Concentration in Humans: data-driven glucose modeling," vol. 56, no. 2, pp. 246–254, 2009.
- [142] A. Gani, A. V. Gribok, Y. Lu, W. K. Ward, R. a. Vigersky, and J. Reifman, "Universal glucose models for predicting subcutaneous glucose concentration in humans," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 1, pp. 157–165, 2010.
- [143] H. Kirchsteiger, R. Johansson, E. Renard, and L. D. Re, "Continuous-time interval model identification of blood glucose dynamics for type 1 diabetes," *International Journal of Control*, vol. 87, no. 7, pp. 1454–1466, 2014.

- [144] N. A. Khovanova, I. A. Khovanov, L. Sbrano, F. Griffiths, and T. A. Holt, "Characterization of Linear Predictability and Non-stationarity of Subcutaneous Glucose Profiles," *Computer Methods and Programs in Biomedicine*, vol. 110, no. 3, pp. 260–267, 2013.
- [145] R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*, vol. 29. 1994.
- [146] F. Q. Nuttall, M. C. Gannon, J. L. Wald, and M. Ahmed, "Plasma glucose and insulin profiles in normal subjects ingesting diets of varying carbohydrate, fat, and protein content," *Journal of the American College of Nutrition*, vol. 4, no. 4, pp. 437–450, 1985.
- [147] P. Stahel, J. P. Cant, J. a. R. MacPherson, H. Berends, and M. a. Steele, "A mechanistic model of intermittent gastric emptying and Glucose-Insulin dynamics following a meal containing milk components," *PLoS ONE*, vol. 11, no. 6, pp. 1–17, 2016.
- [148] M. Cescon, R. Johansson, E. Renard, and A. Maran, "Identification of individualised empirical models of carbohydrate and insulin effects on T1DM blood glucose dynamics," *International Journal of Control*, vol. 87, no. 7, pp. 1438–1453, 2014.
- [149] I. M. Gel'fand and G. E. Shilov, *Generalized functions. Vol. I: properties and operations*. Academic Press, New York, 1964.
- [150] S. H. Strogatz, *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Westview press, 2014.
- [151] J. A. Siegel, J. L. Urbain, L. P. Adler, N. D. Charkes, A. H. Maurer, B. Krevsky, L. C. Knight, R. S. Fisher, and L. S. Malmud, "Biphasic nature of gastric emptying," *Gut*, vol. 29, no. 1, pp. 85–89, 1988.
- [152] Gonlachvanit S, Hsu C-W, and Boden GH, "Effect of altering gastric emptying on postprandial plasma glucose concentrations following a physiologic meal in type-II diabetic patients," *Digestive Diseases and Sciences*, vol. 48, no. 3, pp. 188–197, 2003.
- [153] D. J. Michael, R. A. Ritzel, L. Haataja, and R. H. Chow, "Slow-release forms," *Diabetes*, vol. 55, no. March, pp. 600–607, 2006.

- [154] R. S. Parker, F. J. Doyle III, J. H. Ward, and N. A. Peppas, "Robust H glucose control in diabetes using a physiological model," *AIChE Journal*, vol. 46, no. 12, pp. 2537–2549, 2000.
- [155] C. Mathworks, "SimBiology ® Release Notes."
- [156] R. Higgins, D. Lowe, M. Hathaway, H. Kashi, L. C. Tan, C. Imray, S. Fletcher, K. Chen, N. Krishnan, R. Hamer, and Others, "Rises and falls in donor-specific and third-party HLA antibody levels after antibody incompatible transplantation," *Transplantation*, vol. 87, no. 6, pp. 882–888, 2009.
- [157] Y. Zhang, D. Lowe, D. Briggs, R. Higgins, and N. Khovanova, "Novel data-driven stochastic model for antibody dynamics in kidney transplantation," *IFAC-PapersOnLine*, vol. 48, no. 20, pp. 249–254, 2015.
- [158] Y. Zhang, D. Briggs, D. Lowe, D. Mitchell, S. Daga, N. Krishnan, R. Higgins, and N. Khovanova, "A new data-driven model for post-transplant antibody dynamics in high risk kidney transplantation- in press," *Mathematical Biosciences*, pp. 1–9, 2016.
- [159] J. Gloor and M. D. Stegall, "Sensitized renal transplant recipients: current protocols and future directions," *Nature Reviews Nephrology*, vol. 6, no. 5, pp. 297–306, 2010.
- [160] H. D. Nguyen, R. L. Williams, G. Wong, and W. H. Lim, "The evolution of HLA-matching in kidney transplantation," in *Current Issues and Future Direction in Kidney Transplantation*, INTECH Open Access Publisher, 2013.
- [161] J. Robinson, J. A. Halliwell, H. McWilliam, R. Lopez, and S. G. E. Marsh, "IPD - The Immuno Polymorphism Database," *Nucleic Acids Research*, vol. 41, no. D1, pp. 1234–1240, 2013.
- [162] D. C. Brennan, "HLA matching and graft survival in kidney transplantation," *UpToDate. Waltham, MA: UpToDate*, 2012.
- [163] I. I. N. Doxiadis, J. Smits, G. M. Th Schreuder, G. G. Persijn, H. C. van Houwelingen, J. J. van Rood, and F. H. J. Claas, "Association between specific HLA combinations and probability of kidney allograft loss: the taboo concept," *Lancet*, vol. 348, no. 9031, pp. 850–853, 1996.

- [164] A. V. Reisæ ter, T. Leivestad, F. Vartdal, A. Spurkland, P. Fauchald, I. B. Brekke, and E. Thorsby, “A strong impact of matching for a limited number of HLA-DR antigens on graft survival and rejection episodes: a single-center study of first cadaveric kidneys to nonsensitized recipients,” *Transplantation*, vol. 66, no. 4, p. 523, 1998.
- [165] S. Takemoto, F. K. Port, F. H. J. Claas, and R. J. Duquesnoy, “HLA Matching for Kidney Transplantation,” *Human immunology*, vol. 65, pp. 1489–1505, Dec. 2004.
- [166] S. K. Takemoto, P. I. Terasaki, D. W. Guertson, and J. M. Cecka, “Twelve years’ experience with national sharing of HLA-matched cadaveric kidneys for transplantation,” *The New England Journal of Medicine*, vol. 343, no. 15, pp. 1078–1084, 2000.
- [167] L. D. Cornell, R. N. Smith, and L. B. Colvin, “Kidney Transplantation: Mechanisms of Rejection and Acceptance,” *Annual Review of Pathology-mechanisms of Disease*, vol. 3, no. 2, pp. 189–220, 2008.
- [168] M. Cascalho and J. L. Platt, “Basic mechanisms of humoral rejection,” *Pediatric Transplantation*, vol. 9, no. 1, pp. 9–16, 2005.
- [169] S. B. Moore, S. Sterioff, A. M. Pierides, S. K. Watts, and C. M. Ruud, “Transfusion-induced alloimmunization in patients awaiting renal allografts,” *Vox sanguinis*, vol. 47, no. 5, pp. 354–361, 1984.
- [170] P. C. Lee, P. I. Terasaki, S. K. Takemoto, P. H. Lee, C. J. Hung, Y. L. Chen, A. Tsai, and H. Y. Lei, “All chronic rejection failures of kidney transplants were preceded by the development of HLA antibodies,” *Transplantation*, vol. 74, no. 8, pp. 1192–1194, 2002.
- [171] J. E. Worthington, S. Martin, D. M. Al-Husseini, P. A. Dyer, and R. W. G. Johnson, “Posttransplantation production of donor HLA-specific antibodies as a predictor of renal transplant outcome,” *Transplantation*, vol. 75, no. 7, pp. 1034–1040, 2003.
- [172] P. I. Terasaki and J. Cai, “Human leukocyte antigen antibodies and chronic rejection: from association to causation,” *Transplantation*, vol. 86, pp. 377–383, Aug. 2008.

- [173] A. Picascia, T. Infante, and C. Napoli, "Luminex and antibody detection in kidney transplantation," *Clinical and Experimental Nephrology*, vol. 16, no. 3, pp. 373–381, 2012.
- [174] The British Transplantation Society, "Guidelines for antibody incompatible transplantation," 2011.
- [175] N. S. Krishnan, D. Zehnder, D. Briggs, and R. Higgins, "Human leukocyte antigen antibody incompatible renal transplantation," *Indian journal of nephrology*, vol. 22, no. 6, p. 409, 2012.
- [176] J. Fotheringham, C. A. Angel, and W. McKane, "Transplant glomerulopathy: morphology, associations and mechanism," *Nephron Clinical Practice*, vol. 113, no. 1, pp. c1–c7, 2009.
- [177] J. Sellarés, D. G. De Freitas, M. Mengel, J. Reeve, G. Einecke, B. Sis, L. G. Hidalgo, K. Famulski, A. Matas, and P. F. Halloran, "Understanding the causes of kidney transplant failure: the dominant role of antibody-mediated rejection and nonadherence," *American Journal of Transplantation*, vol. 12, no. 2, pp. 388–399, 2012.
- [178] A. A. Zachary and M. S. Leffell, "Detecting and monitoring human leukocyte antigen-specific antibodies," *Human immunology*, vol. 69, no. 10, pp. 591–604, 2008.
- [179] C. Puttarajappa, R. Shapiro, and H. P. Tan, "Antibody-mediated rejection in kidney transplantation: a review," *Journal of transplantation*, vol. 2012, pp. 1–9, Jan. 2012.
- [180] J. E. Worthington, A. McEwen, L. J. McWilliam, M. L. Picton, and S. Martin, "Association between C4d staining in renal transplant biopsies, production of donor-specific HLA antibodies, and graft outcome," *Transplantation*, vol. 83, no. 4, pp. 398–403, 2007.
- [181] P. I. Terasaki and M. Ozawa, "Predicting kidney graft failure by HLA antibodies: a prospective trial," *American Journal of Transplantation*, vol. 4, pp. 438–443, Mar. 2004.

- [182] P. Terasaki and K. Mizutani, "Antibody mediated rejection: update 2006," *Clinical journal of the American Society of Nephrology*, vol. 1, no. 3, pp. 400–403, 2006.
- [183] L. G. Hidalgo, P. M. Campbell, B. Sis, G. Einecke, M. Mengel, J. Chang, J. Sellares, J. Reeve, and P. F. Halloran, "De novo donor-specific antibody at the time of kidney transplant biopsy associates with microvascular pathology and late graft failure," *American Journal of Transplantation*, vol. 9, no. 11, pp. 2532–2541, 2009.
- [184] C. Wiebe, I. W. Gibson, T. D. Blydt-Hansen, M. Karpinski, J. Ho, L. J. Storsley, A. Goldberg, P. E. Birk, D. N. Rush, and P. W. Nickerson, "Evolution and clinical pathologic correlations of de novo donor-specific HLA antibody post kidney transplant," *American Journal of Transplantation*, vol. 12, no. 5, pp. 1157–1167, 2012.
- [185] M. Willicombe, P. Brookes, E. Santos-Nunez, J. Galliford, A. Ballow, A. Mclean, C. Roufosse, H. T. Cook, A. Dorling, a. N. Warrens, T. Cairns, and D. Taube, "Outcome of patients with preformed donor-specific antibodies following alemtuzumab induction and tacrolimus monotherapy," *American Journal of Transplantation*, vol. 11, no. 3, pp. 470–477, 2011.
- [186] R. M. Higgins, D. J. Bevan, B. S. Carey, C. K. Lea, M. Fallon, R. Bühler, R. W. Vaughan, P. J. O'Donnell, S. A. Snowden, M. Bewick, and B. M. Hendry, "Prevention of hyperacute rejection by removal of antibodies to HLA immediately before renal transplantation," *Lancet*, vol. 384, pp. 1208–1211, Nov. 1996.
- [187] R. Higgins, D. Lowe, M. Hathaway, F. T. Lam, H. Kashi, L. C. Tan, C. Imray, S. Fletcher, K. Chen, N. Krishnan, R. Hamer, D. Zehnder, and D. Briggs, "Double filtration plasmapheresis in antibody-incompatible kidney transplantation," *Therapeutic Apheresis and Dialysis*, vol. 14, no. 4, pp. 392–399, 2010.
- [188] N. S. Krishnan, D. Zehnder, S. Daga, D. Lowe, F. T. Lam, H. Kashi, L. C. Tan, C. Imray, R. Hamer, D. Briggs, N. Raymond, and R. M. Higgins, "Behaviour of Non-donor sSpecific antibodies during rapid re-synthesis of donor specific HLA antibodies after antibody incompatible renal transplantation," *PLoS ONE*, vol. 8, no. 7, p. e68663, 2013.

- [189] P. S. de Souza, E. David-Neto, N. Panajotopolous, F. Agena, H. Rodrigues, C. Ronda, D. R. David, J. Kalil, W. C. Nahas, and M. C. R. de Castro, "Dynamics of anti-human leukocyte antigen antibodies after renal transplantation and their impact on graft outcome," *Clinical Transplantation*, vol. 28, no. 11, pp. 1234–1243, 2014.
- [190] R. Higgins, M. Hathaway, D. Lowe, H. Kashi, L. C. Tan, C. Imray, S. Fletcher, D. Zehnder, K. Chen, and N. Krishnan, "Blood levels of donor-specific human leukocyte antigen antibodies after renal transplantation: resolution of rejection in the presence of circulating donor-specific antibody," *Transplantation*, vol. 84, no. 7, pp. 876–884, 2007.
- [191] R. Higgins, D. Lowe, M. Hathaway, C. Williams, H. Kashi, L. C. Tan, C. Imray, S. Fletcher, K. Chen, N. Krishnan, R. Hamer, S. Daga, M. Edey, D. Zehnder, and D. Briggs, "Human leukocyte antigen antibody-incompatible renal transplantation: excellent medium-term outcomes with negative cytotoxic crossmatch," *Transplantation*, vol. 92, no. 8, pp. 900–906, 2011.
- [192] A. Morell, W. D. Terry, and T. A. Waldmann, "Metabolic properties of IgG subclasses in man," *Journal of Clinical Investigation*, vol. 49, no. 4, pp. 673–680, 1970.
- [193] E. Reed, M. Hardy, A. Benvenisty, C. Lattes, J. Brensilver, R. McCabe, K. Reemstma, D. W. King, and N. Suciu-Foca, "Effect of antiidiotypic antibodies to HLA on graft survival in renal-allograft recipients," *New England Journal of Medicine*, vol. 316, no. 23, pp. 1450–1455, 1987.
- [194] R. Pei, J. how Lee, N. J. Shih, M. Chen, and P. I. Terasaki, "Single human leukocyte antigen flow cytometry beads for accurate identification of human leukocyte antigen antibody specificities," *Transplantation*, vol. 75, no. 1, pp. 43–49, 2003.
- [195] C. Lefaucheur, A. Loupy, G. S. Hill, J. Andrade, D. Nochy, C. Antoine, C. Gautreau, D. Charron, D. Glotz, and C. Suberbielle-Boissel, "Preexisting donor-specific HLA antibodies predict outcome in kidney transplantation," *Journal of the American Society of Nephrology*, vol. 21, pp. 1398–1406, Aug. 2010.
- [196] O. Thaunat, W. Hanf, V. Dubois, B. McGregor, G. Perrat, C. Chauvet, J.-L. Touraine, and E. Morelon, "Chronic humoral rejection mediated by anti-HLA-DP

- alloantibodies: insights into the role of epitope sharing in donor-specific and non-donor specific alloantibodies generation,” *Transplant immunology*, vol. 20, pp. 209–211, Mar. 2009.
- [197] E. S. Woodle, a. R. Shields, N. S. Ejaz, B. Sadaka, A. Girnita, R. C. Walsh, R. R. Alloway, P. Brailey, M. a. Cardi, B. G. Abu Jawdeh, P. Roy-Chaudhury, A. Govil, and G. Mogilishetty, “Prospective iterative trial of proteasome inhibitor-based desensitization,” *American Journal of Transplantation*, vol. 15, no. 1, pp. 101–118, 2015.
- [198] MATLAB, *Version 8.2.0.701 (2013b)*. The MathWorks Inc., 2013.
- [199] L. Ljung, “System identification toolbox for use with MATLAB,” 2007.
- [200] G. C. Cawley and N. L. C. Talbot, “Fast exact leave-one-out cross-validation of sparse least-squares support vector machines,” *Neural networks : the official journal of the International Neural Network Society*, vol. 17, pp. 1467–1475, Dec. 2004.
- [201] E. F. Reed, P. Rao, Z. Zhang, H. Gebel, R. A. Bray, I. Guleria, J. Lunz, T. Mohanakumar, P. Nickerson, A. R. Tambur, A. Zeevi, P. S. Heeger, and D. Gjertson, “Comprehensive assessment and standardization of solid phase multiplex-bead arrays for the detection of antibodies to HLA,” *American Journal of Transplantation*, vol. 13, no. 7, pp. 1859–1870, 2013.
- [202] L. Arnold, *Stochastic Differential Equations : Theory and Applications*. Krieger Publishing Company, 1973.
- [203] N. Khovanova, S. Daga, T. Shaikhina, N. Krishnan, J. Jones, D. Zehnder, D. Mitchell, R. Higgins, D. Briggs, and D. Lowe, “Subclass analysis of donor HLA-specific IgG in antibody-incompatible renal transplantation reveals a significant association of IgG 4 with rejection and graft failure,” *Transplant International*, vol. 28, pp. 1405–1415, 2015.
- [204] M. J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, and D. L. Wild, “A Bayesian approach to reconstructing genetic regulatory networks with hidden factors,” *Bioinformatics*, vol. 21, no. 3, pp. 349–356, 2005.

-
- [205] C. Y. Y. Lee and M. P. Wand, “Variational methods for fitting complex Bayesian mixed effects models to health data,” *Statistics in Medicine*, vol. 35, no. 2, pp. 165–188, 2016.

Appendix: Matlab code

As introduced in Chapter 2, a VB toolbox [77], initially developed for neuroimaging data, was applied to perform the inference in this thesis. It can be accessed via <http://mbb-team.github.io/VBA-toolbox>. The specific codes for the models that have been built in Chapters 2, 3, and 4 are shown as follows:

```
1 %%Chapter 2 f_chp2 is the model for the example shown in Section 2.8
2 function [fx,dF_dX,dF_dTheta,d2F_dXdTheta] = f_chp2(Xt,Theta,ut,inF)
3 % function [f,J] = f_doubleWell(t,x,theta)
4 %
5 % This function computes the evolution function that comes from a .
6 % IN:
7 %   - t: time index (not used here)
8 %   - x: the current state of the system
9 %   - theta: a 2x1 vector parameter (containing the position of the two
10 %     wells.
11 % OUT:
12 %   - f: the current value of the evolution function
13 %   - J: the jacobian of the system
14
15 deltata = inF.deltat;
16 k       = Theta(1);
17 b       = Theta(2);
18 a       = Theta(3);
19 x1      = Xt(1);
20 x2      = Xt(2);
21
22 f       = [ x2 ; -a-b*x1-k*x2 ];
23 J       = [ 0      1
24            -b      -k ];
25
26 fx      = deltata.*f + Xt;
27 dF_dX   = deltata.*J' + eye(2);
28 dF_dTheta = deltata.*[ 0 -x2
29                       0 -x1
30                       0 -1];
```

```

31 d2F_dXdTheta(:, :, 1) = zeros(2, 3);
32 d2F_dXdTheta(:, :, 2) = deltat.*[ 0 -1 0
33                                     -1 0 0];
34
35
36 %% f_ML and f_M2 are the models ML and M2 described in Chapter 3
37 function [fx, dF_dX, dF_dTheta, d2F_dXdTheta] = f_ML(Xt, Theta, ut, inF)
38 deltat = inF.deltat;
39 k = Theta(1);
40 b = Theta(2);
41 x1 = Xt(1);
42 x2 = Xt(2);
43
44 f = [ x2 ; -b*x1-k*x2 ];
45 J = [ 0 1
46       -b -k ];
47
48 fx = deltat.* f + Xt;
49 dF_dX = deltat.*J' + eye(2);
50 dF_dTheta = deltat.*[ 0 -x2
51                       0 -x1];
52 d2F_dXdTheta(:, :, 1) = zeros(2, 2);
53 d2F_dXdTheta(:, :, 2) = deltat.*[ 0 -1
54                                     -1 0];
55
56
57 function [fx, dF_dX, dF_dTheta, d2F_dXdTheta] = f_M2(Xt, Theta, ut, inF)
58 deltat = inF.deltat;
59 k0 = Theta(1);
60 k1 = Theta(2);
61 k2 = Theta(3);
62 b = Theta(4);
63 x1 = Xt(1);
64 x2 = Xt(2);
65
66 f = [ x2 ; -b*x1-k0*x2-k1*x1*x2-k2*x1^2*x2 ];
67 J = [ 0 1
68       -b-k1*x2-2*k2*x1*x2 -k0-k1*x1-k2*x1^2 ];
69
70 fx = deltat.* f + Xt;
71 dF_dX = deltat.*J' + eye(2);
72 dF_dTheta = deltat.*[ 0 -x2
73                       0 -x1*x2
74                       0 -x1^2*x2
75                       0 -x1];
76 d2F_dXdTheta(:, :, 1) = zeros(2, 4);
77 d2F_dXdTheta(:, :, 2) = deltat.*[ 0 -x2 -2*x1*x2 -1
78                                     -1 -x1 -x1^2 0];

```

```

79
80
81 %% f_chp4 is the model described in Chapter 4.3
82 function [fx,dF_dX,dF_dTheta,d2F_dXdTheta] = f_chp4(Xt,Theta,ut,inF)
83 deltat = inF.deltat;
84
85 a = Theta(1);
86 b = Theta(2);
87 c = Theta(3);
88 d = Theta(4);
89 x1 = Xt(1);
90 x2 = Xt(2);
91 x3 = Xt(3);
92
93 f = [ x2 ; x3; -a*x1-b*x2-c*x3-d ];
94 J = [ 0      1      0
95       0      0      1
96      -a     -b     -c ];
97
98 fx = deltat.*(f+ut*[0;1;0]) + Xt;
99 dF_dX = deltat.*J' + eye(3);
100 dF_dTheta = deltat.*[ 0 0 -x1
101                       0 0 -x2
102                       0 0 -x3
103                       0 0 -1];
104 d2F_dXdTheta(:,:,1) = zeros(3,4);
105 d2F_dXdTheta(:,:,2) = zeros(3,4);
106 d2F_dXdTheta(:,:,3) = deltat.*[ -1 0 0 0
107                                   0 -1 0 0
108                                   0 0 -1 0];

```
